Understanding the Mechanisms of Robustness in Intracellular Protein Signalling Cascades and Gene Expression

Von der Fakultät Konstruktions-, Produktions- und Fahrzeugtechnik
und dem Stuttgart Research Centre for Simulation Technology
der Universität Stuttgart zur Erlangung der Würde eines
Doktor-Ingenieurs (Dr.-Ing.) genehmigte Abhandlung

Vorgelegt von

D. Paul

aus Howrah, Indien

Hauptberichter:	Prof. Dr. rer. nat. Nicole Radde
Mitberichter:	Prof. Dr.-Ing. habil. Manfred Bischoff
	Prof. Jeremy Gunawardena, PhD (Mathematics)

Tag der mündlichen Prüfung: 30. April 2019

Institut für Systemtheorie und Regelungstechnik

Universität Stuttgart

2018

To my parents & professors

ā no bhadrāḥ kratavo kṣyantu viśvato

(*Let noble thoughts come to me from all directions*)

– *Rigveda* 1.89.1

Acknowledgements

During the culmination of a thesis, a sense emerges that acknowledgements are due, and that is an opportunity to permanently document one's gratitude to people who are special and without whose support this point would be more difficult to reach. Therefore, I would like to express my sincere gratitude and regards to my supervisor, **Prof. Dr. rer. nat. Nicole Radde**, for her constant cooperation, encouragement, rationalistic criticism and guidance during the entire duration of my doctorate.

I am deeply thankful to my collaborator and advisor **Prof. Dr.-Ing. habil. Manfred Bischoff**, and my project partner **M.Sc. Layla Koohi Fayegh Dehkordi** from the Institute for Structural Mechanics of University of Stuttgart. I am truely blessed with the opportunity to carry out a part of my doctoral research for three months in **Prof. Jeremy Gunawardena**'s group at the Harvard Medical School, USA. In this respect, I would also like to thank **Chris Nam** from the Gunawardena lab for being my project collaborator.

I am extremely delighted to have wonderful colleagues like Antje, Caterina, Dirke, Wolfgang, Raffaele. I would like to thank all of them for their help, support and encouragements during my stay at the institute. I am especially grateful to Antje, Wolfgang, Dirke, Raffaele, and Simon Niederländer for lending their valuable time to proof-read the thesis. In addition, I would like to thank Dr. Victoria Grushkovskaya and Dr. Zoltan Tuza, for a wonderful time together at the institute.

My doctoral journey would not have been smooth without the secretaries of the institute. Therefore, my sincere thanks go to Sabine Balschat, Beate Spinner, Claudia Vetter, and most importantly Norvi Brendle-Forero for their patient and efficient handling of complicated bureaucratic matters during my doctoral studies.

Thank you Caterina and Francesco, Simona, Alicia, Marc, and Anna for being my friends and family away from home.

Finally, I would like to acknowledge German Research Foundation (DFG) as part of the Transregional Collaborative Research Centre (SFB/Transregio) 141 'Biological Design and Integrative Structures'/project B05; and the Cluster of Excellence in Simulation Technology (EXC 310/2) at the University of Stuttgart for funding my doctoral research and my research stay at the Harvard Medical School, USA, respectively.

Stuttgart, July 2019
Debdas Paul

Contents

Summary

We seek to understand the structural as well as the mechanistic basis of robustness in intracellular protein signalling cascades and in transcriptional regulation of gene expression. For protein signalling cascades, we employ a comparison based study involving a single, a double and a cascade of two double phosphorylation-dephosphorylation (PD) cycles. Using deterministic modelling approaches based on ordinary differential equations (ODE), we observe that the cascade of two double PD cycles exhibits robust output behaviour compared to that of a single and a double PD cycle upon constant as well as time-varying input perturbations. Furthermore, a system theoretic analysis reveals that the protein phosphorylation cascades act as an efficient low-pass filter that attenuates the noise mimicked as high-frequency input signals. Afterwards, we extend the study for a stochastic environment. Simulation results based on the stochastic simulation algorithm (SSA) reveal a novel phenomenon called dynamic sequestration that plays an ambivalent role as an intrinsic noise filter. Overall, the analysis indicates that complexity can be one of the basic principles of robust biological designs such as intracellular protein signalling cascades.

A major function of intracellular signalling cascades is to transmit the extracellular signal to the nucleus to initiate the process of gene expression. Gene expression is an intrinsically stochastic process that results into cell-to-cell variability in protein and messenger RNA (mRNA) levels, often termed as the expression noise. In spite of such noise, how cells achieve robustness is therefore a fundamental biological problem. We conclude the thesis by introducing a rule-based modelling approach based on the Kappa (κ) platform with the goal to understand the underlying mechanisms that ensure robust cellular functioning during gene expression. In particular, we introduce a gene expression model that keeps the process of transcription and excludes the process of translation. Therefore, we quantify the expression noise using mRNA which is the end product of transcription. Besides, the motivation behind adopting a rule-based modelling approach is that unlike the ODE-based approach, the former subsumes the combinatorial complexity arises due to various binding configurations of transcription factors (TF) for regulation of gene expression and offers a compact graphical representation of the same. Afterwards, the representation is transformed into an equivalent set of executable κ rules that are simulated using the SSA to obtain distributions of mRNA copy numbers corresponding to different regulatory mechanisms.

Deutsche Zusammenfassung

Verständnis der Mechanismen der Robustheit bei intrazellulären Proteinsignalkaskaden und Genexpressionen

Wir wollen sowohl die strukturellen als auch die mechanistischen Grundlagen der Robustheit in intrazellulären Proteinsignalkaskaden und in der transkriptionellen Regulation der Genexpression verstehen. Für die Untersuchung von Proteinsignalkaskaden verwenden wir eine vergleichsbasierte Studie mit einer Einzelphosphorylierung, einer Doppelphosphorylierung und einer Kaskade von zwei Doppelphosphorylierungs-Dephosphorylierungs-(PD)-Zyklen. Zur Modellierung verwenden wir deterministische Ansätze, die auf gewöhnlichen Differentialgleichungen (ODE) basieren. Im Gegensatz zu einem einzelnen und einem doppelten PD-Zyklus weist die Kaskade von zwei doppelten PD-Zyklen ein robustes Ausgabeverhalten bei konstanten sowie zeitvariablen Eingangsstörungen auf. Darüber hinaus zeigt eine systemtheoretische Analyse, dass die Proteinphosphorylierungskaskaden als effizienter Tiefpassfilter wirken, der hochfrequente Eingangssignale dämpft. Anschließend erweitern wir die Studie mit einer stochastischen Umgebung. Simulationsergebnisse, die auf dem stochastischen Simulationsalgorithmus (SSA) basieren, zeigen ein neuartiges Phänomen namens "Dynamic Sequestration", das eine ambivalente Rolle als intrinsischer Rauschfilter spielt. Insgesamt zeigt die Analyse, dass Komplexität eines der Grundprinzipien robuster biologischer Systeme wie intrazellulärer Proteinsignalkaskaden sein kann.

Eine der Hauptfunktionen intrazellulärer Signalkaskaden besteht darin das extrazelluläre Signal an den Kern zu übertragen, um den Prozess der Genexpression einzuleiten. Die Genexpression ist ein intrinsisch stochastischer Prozess, der zu einer Variabilität der Protein- und Messenger-RNA (mRNA)-Menge von Zelle zu Zelle führt, die oft als Expressionsrauschen bezeichnet wird. Trotz des Rauschens ist es daher ein grundlegendes biologisches Problem, wie Zellen ihre Robustheit erreichen. Um zugrunde liegende Mechanismen zu

verstehen, die eine robuste zelluläre Funktion während der Genexpression gewährleisten, schließen wir die Arbeit mit der Einüfhrung eines regelbasierten Modellierungsansatzes auf Basis der Kappa (κ)-Plattform ab. Insbesondere stellen wir ein Genexpressionsmodell vor, das den Prozess der Transkription beibehält und den Prozess der Translation ausschließt. Daher quantifizieren wir das Expressionsrauschen mit Hilfe der mRNA, die das Endprodukt der Transkription ist. Darüber hinaus ist die Motivation für die Verwendung eines regelbasierten Modellierungsansatzes, dass im Gegensatz zum ODE-basierten Ansatz die kombinatorische Komplexität durch verschiedene Bindungskonfigurationen von Transkriptionsfaktoren (TF) zur Regulierung der Genexpression abgebildet wird und eine kompakte grafische Darstellung derselben geboten wird. Anschließend wird die Darstellung in einen äquivalenten Satz von ausführbaren κ-Regeln umgewandelt, die mit Hilfe der SSA simuliert werden, um Verteilungen von mRNA-Molekülen zu erhalten, die verschiedenen Regulationsmechanismen entsprechen.

Nomenclature

The following notational conventions are maintained throughout the entire thesis.

Sets of numbers

Symbol	Description
$\mathbb{N}$	natural numbers
$\mathbb{Z}$	integers
$\mathbb{Z}_{>0}$	positive integers
$\mathbb{R}$	real numbers
$\mathbb{R}_{\geq 0}$	non-negative real numbers
$\mathbb{R}_{>0}$	positive real numbers

Statistics

Symbol	Description
$\mathrm{Var}[\delta]$	variance of δ
$\mathbf{E}[\delta]$	expectation of δ

Variables

Symbol	Description
u	upstream input signal (constant or time varying) in a phosphorylation-dephosphorylation cascade
$\bar{x}_{\mathrm{M}}$	steady state output ($\bar{x}$) of protein X in model M
s^{M}	normalized local sensitivity coefficients for the steady state output of model M
f	frequency
$A^*(f)$	output semi-amplitude at input frequency f

c_v^{ss}	coefficient of variation of the model output (ppX or ppY) at steady state
c_v^{M}	coefficient of variation of the steady state output for model M
ρ	Pearson's correlation coefficient
r_s	Spearman's correlation coefficient
Γ,γ	autocovariance and autocorrelation functions
X, pX, ppX	protein X, it's singly, and doubly phosphorylated form
$\underbrace{\mathrm{p}\ldots\mathrm{p}}_{m}\mathrm{X}$	$m \in \mathbb{Z}_{>0}$ times phosphorylated form of protein X
$\mathrm{E}_{\mathrm{kin}}$	kinase molecule
$\mathrm{E}_{\mathrm{pho}}$	phosphatase molecule
k_{on}, k_{off}, and k_{cat}	Binding, unbinding and catalytic rate constants for a two step enzyme-substrate kinetics (see Section 3.2 for more details)
$\mathrm{B}_{k^{\mathrm{X}}=k^{\mathrm{Y}}}$	k_{on}, k_{off} and k_{cat} values of the X protein and the Y protein modules are pairwise equal
$\mathrm{B}_{k^{\mathrm{Y}}_{\{\mathrm{off,\ cat}\}}=r*k^{\mathrm{X}}_{\{\mathrm{off,\ cat}\}}}$	k_{off} and k_{cat} of the Y protein module are $r \in \mathbb{R}_{>0}$ times that of the X protein module
ppX^t	total number (including complexes) of doubly phosphorylated molecules of X protein
$\mathrm{Y}^t + \mathrm{pY}^t$	total number of unphosphorylated and singly phosphorylated molecules of Y protein
Y^T	total number Y protein molecules, considered as a conserved quantity
$\overline{\mathrm{ppX}}$	number of unbounded ppX molecules in steady state
$\overline{\mathrm{ppX}}^t$	total number of ppX molecules in steady state

Operators

Symbol	**Description**
$\nabla_\gamma \Psi(\gamma, \cdots)$	differentiation of Ψ with respect to γ
$H(P, Q)$	Hamming distance between two binary vectors P and Q

Physical constants

Symbol	Description	Value
A_v	Avogadro constant	$6.022140857(74) \times 10^{23}$ mol^{-1}
k_B	the Boltzmann constant	$1.38064852(79) \times 10^{-23}$ J·K^{-1}

Logical symbols

Symbol	Description
$\wedge$	logical AND
$\vee$	logical OR
$\sim$	logical NOT

Acronyms

Signalling pathways

MAPK	Mitogen-activated protein kinase
ERK	extracellular signal-regulated kinase

Techniques

smFISH	single-molecule fluorescence *in situ* hybridization

Abbreviations

Cellular and molecular biology

PD	phosphoryation-dephosphorlyation
PTMs	post-translational modifications
ATP	Adenosine triphosphate
DNA	Deoxyribonuclic acid
RNA	Ribonuclic acid
mRNA	messenger RNA
ASF/SF2	alternative splicing factor/pre-mRNA-splicing factor

SRPK1	Serine/threonine-protein kinase 1
TF	transcription factor
RNAP	RNA polymerase
RNAPI, RNAPII, and RNAPIII	RNA polymerase I, RNA polymerase II, and RNA polymerase III
PIC	pre-initiation complex consists of TFs and RNAPII
DRB	5,6-Dichlorobenzimidazole 1-β-D-Ribofuranoside
DSIF	DRB sensitivity inducing factor
NELF	negative elongation factor
p-TEFb	positive transcription elongation factor
TFIIS	transcription elongation factor IIS
nt	nucleotide

Mathematical equations

CME	chemical master equation
ODE	ordinary differential equation

Statistics

CV	coefficient of variation

Algorithm

SSA	stochastic simulation algorithm

Systems theory

LTI	linear time-invariant system

Stochastic model

CTMC	Continuous-time Markov chains

1 Introduction

The term *robustness* provides an intuitive idea about failure or sustainability of a biological or technical system under uncertainties and rapid changes in the environment of that system. According to Alderson and Doyle, the definition of robustness has multiple aspects and depends on the *properties* of the system. For example, *reliability* is the robustness against failure of a component, *efficiency* is the robustness against the scarcity of resources, and modularity ensures robustness to recognize unit interactions (Alderson and Doyle, 2010). It is not necessary that a property or feature of the system is always quantifiable. For example, robust design of auction against collusion is not a quantifiable property (Jen, 2005)[1]. This is one of the critical aspects where the concept of robustness differs from the concept of an analogous term called *stability*. Stability theory has a precise mathematical formulation, and hence provides a quantitative measure of the *persistence* of the system's properties under perturbations. Robustness, in addition, leads to a deeper understanding of the system's internal properties that are not often easily quantifiable such as evolvability, organization, the interplay between dynamics and organization, multi-functionality, creativity, etc. A detailed discussion on the topic of robustness versus stability is beyond the scope of this thesis. Interested readers are referred to Jen (2005) for a deeper insight.

One of the fundamental challenges in the area of research related to robustness is to understand the underlying *design principle* that leads to robust behaviour of biological and technical systems. Biological systems fundamentally differ from technical systems in two aspects: (1) multi-functionality, and (2) robustness versus optimality trade-off. For technical systems, optimality and robustness are two complementary concepts (Paul et al., 2016). In fact, it is exemplarily shown in Paul et al. (2016) that for technical systems such as load-bearing structures, optimality does not necessarily imply a robust design. On the other hand, biological systems preserve robustness from the perspective of multi-functionality, and at the same time also maintain optimality with respect to specific functions. Bacteria, for example, are incredibly flexible and can maintain functions essential for survival under a variety of conditions (Hart et al., 2011). Thus, unlike technical systems, biological systems accommodate two intrinsically contradictory properties - optimality and robustness. This fact further motivates to investigate the organizational principles of biological systems and

[1] In the field of economics, collusion is an agreement between the parties/firms to limit the open competition between them by dividing the market and setting the price.

their mechanistic aspects in producing robust behaviour. In the next section, we introduce cellular signalling as one of the prime examples of robust biological designs.

1.1 Cellular signalling as a robust biological design

Cellular signalling describes the mechanism through which individual cells *sense* their environment and respond accordingly. Collectively, these sensing and response mechanisms help the organism to dynamically coordinate the activities with the changes in the environment (Krauss, 2006). Apart from adapting to the rapid changes in the external environment, especially higher order organisms perform and maintain various important physiological activities such as cell growth, cell divisions, cell metabolism, etc. through cellular signalling. Biochemically, these physiological activities including the cellular response to the stimuli are facilitated via coordinated work of a group of biomolecules (mainly proteins) forming a chain called the *signalling pathway*. Figure 1.1 provides the big picture of cellular signalling or more precisely intra-cellular signalling (signalling within a cell) where the external signal is transported to the DNA in the cell nucleus through a signalling pathway. The signalling pathway involves a cascade of covalent modification cycles which, in most of the cases, are PD cycles. We call such a formation a *signalling cascade* or a *protein phosphorylation cascade* if the formation involves phosphorylation and dephosphorylation of proteins via kinases and phosphatases. A notable example of such a protein phosphorylation cascade is the Raf-MEK-ERK pathway[2] or collectively the Mitogen-activated protein kinase (MAPK) pathway. In this pathway, first, the singly phosphorylated Raf molecule acts as a kinase to initiate the double phosphorylation of Mitogen-activated protein kinase kinase (MEK) and then the doubly phosphorylated MEK molecule acts as a kinase to facilitate the double phosphorylation of the extracellular signal-regulated kinase (ERK) molecule (Arkun and Yasemi, 2018; Huang and Ferrell, 1996). Figure 1.2 presents a schematic representation of the MAPK pathway where "-p" and "-pp" indicate phosphorylated and doubly phosphorylated forms, respectively. Reliable signalling is indeed essential for survival since malfunctioning compromises fitness for bacteria or perturbs tissue homeostasis, leading to severe diseases. It is known that functional robustness of these pathways is tightly related to the topology of the underlying interaction network and its multi-level regulation (see, e.g., Blüthgen and Legewie (2013); Dexter et al. (2015); Shinar and Feinberg (2010, 2011); Shinar et al. (2009)). Ubiquitous motifs and frequently occurring modules such as PD cycles, contribute to this robustness in several ways, but the exact mechanisms through which robustness is achieved is still unknown for many cases. Therefore, part of this thesis investigates such underlying mechanism in particular for protein phosphorylation cascades

[2] Raf kinase belongs to a family of three serine/threonine kinases (a-Raf, b-Raf, and c-Raf) that are related to retroviral oncognes (Roskoski Jr, 2010)

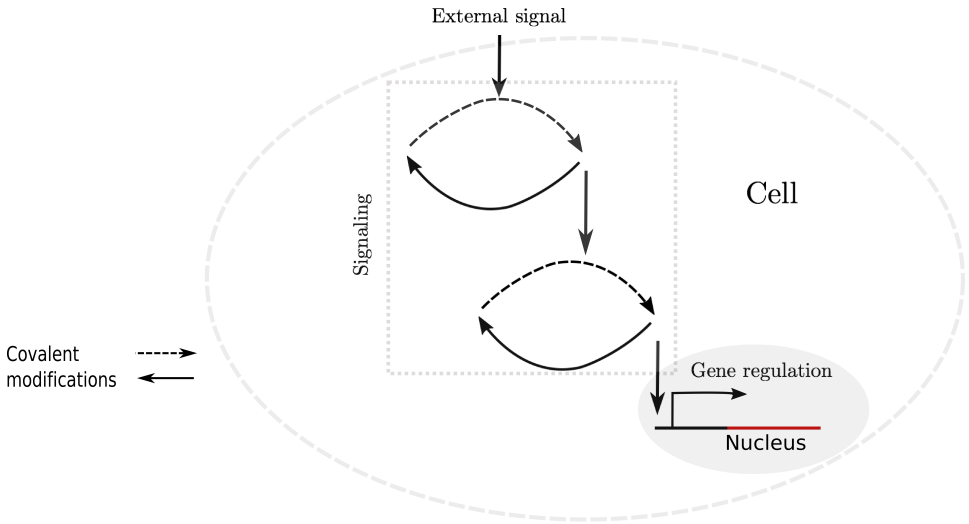


Fig. 1.1. Schematic representation of intracellular signalling. To initiate gene regulation, external signal is carried to the nucleus of the cell through a cascade of covalent modification cycles or PD cycles forming a signalling network motif. In a PD cycle, phosphorylation is indicated by a dashed arrow.

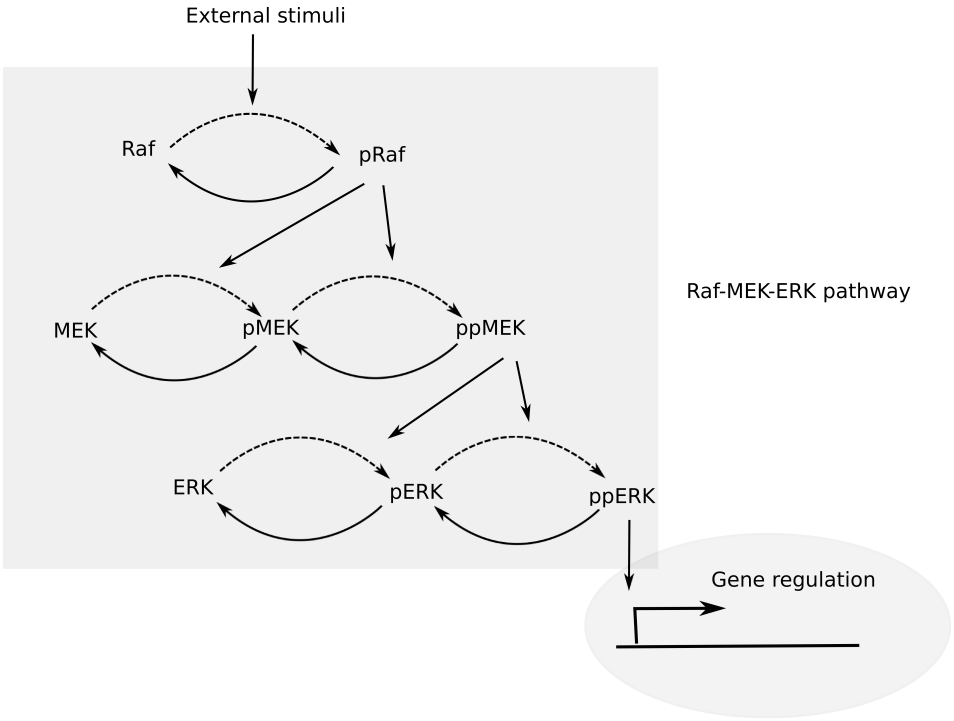


Fig. 1.2. Schematic representation of the Raf-MEK-ERK pathway. The Raf-MEK-ERK pathway receives the external signal from the cell surface and subsequently transmits the signal to the nucleus through a series of PD cycles. The final product of the Raf-MEK-ERK cascade i.e., doubly phosphorylated ERK or activated ERK influence the bindings of TF in the promoter region and thus regulate the gene expression in various ways (Li et al., 2016).

using deterministic and stochastic modelling approaches.

1.2 Gene expression and robustness

The final form of cellular response to its environment comes through the process of gene expression. The process consists of two steps: transcription and translation. During transcription, the double-stranded DNA unwinds and rewinds via the enzyme RNA polymerase (RNAP), and as a result, an mRNA molecule is produced. In the next phase, i.e. during translation, the produced mRNA molecule is decoded in the ribosome to form a chain of amino acids or polypeptides. Later on, the chain folds to an active protein. In this thesis, we focus on the process of eukaryotic transcription that is carried out by RNA polymerase II (RNAPII). Figure 1.3 presents a schematic overview of eukaryotic transcription. A detailed description of the mechanism is provided in Chapter 4.

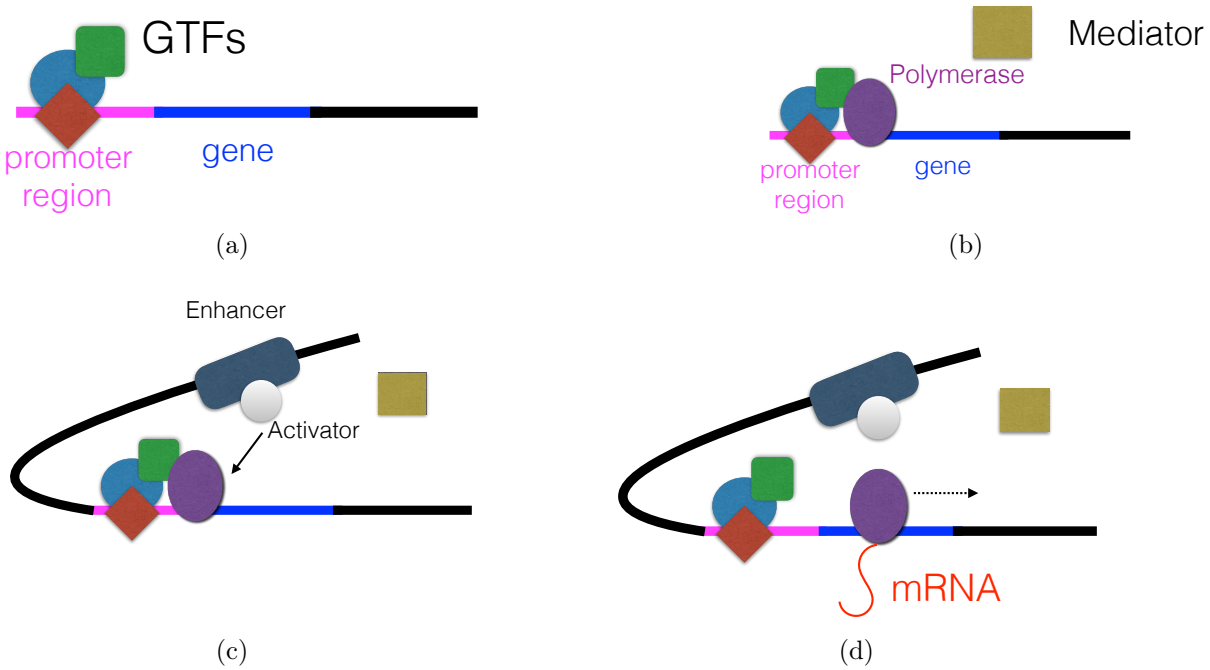


Fig. 1.3. Schematic of eukaryotic gene transcription.(a) General transcription factors (GTFs) are bound to the promoter region to initiate the transcription process. (b) Recruitment of the RNAPII to the promoter region and formation of the preinitiation complex (PIC) together with GTFs. The mediator molecule interacts with the PIC to stabilize it. The RNAPII is now ready to elongate along the DNA, but it can undergo a temporary pause state (c) Enhancer region (50-100 base pairs) of DNA bound to the activator protein forming an enhancer-activator complex that mediates the pause release of RNAPII and increases the likelihood of transcription.(d) Elongation by RNAPII along the DNA, after the release from the paused state, producing nascent mRNA transcripts. The concept is adapted from Kornberg (2007); Myers and Kornberg (2000)

The robustness paradox

Gene expression is inherently stochastic (Raj and van Oudenaarden, 2008). This stochasticity accounts for cell-to-cell variability within the same genotype and environment. The scope of robustness in this context appears to be paradoxical, because robustness implies phenotypic invariance against perturbations, which eventually means reduced variability. As argued in Frank (2007), this paradox arises because an increase in robustness causes the reduction in the adaptivity of a trait or character; hence the selection pressure of evolution is reduced on that trait. As a result of this maladaptation, the performance of the trait is compromised. In a previous communication (Frank, 2003), the author argued that averaging input is a general way to reduce variability in expression phenotypes. Phenotypes which arise by averaging the inputs from many components or cells tend to show robust behaviour for a large number of independent inputs and sample sizes. With larger sample sizes, individual perturbations have less effects on the overall system output (Frank, 2013). Thus, the definition of robustness based on the input-output sensitivity fits in this context. However, the puzzle has not been solved yet. The question now arises how an increase in robustness affects components variability. The answer lies in the way the natural selection acts on phenotypes. Natural selection on the variability of each component weakens with the increase in the number of inputs, thus allowing greater variability in the individual component (Frank, 2013). Expression variability has been found to have a balancing act on the robustness of developmental gene regulation. In an experiment with sea urchin larvae, the authors in Garfield et al. (2013) observed that during the early phase of development expression variation is well buffered; thus regulatory interactions are robust. However, in the later development phase, regulatory interactions become more sensitive to perturbations ensuring variability, increased adaptability for natural selection.

Rule-based model to study the interplay between robustness and stochasticity in gene expression

Study of robustness in gene expression at the level of transcription requires quantification of mRNA copy numbers. An ODE-based approach to this problem suffers from the combinatorics of transcription factor (TF) binding. To alleviate the problem, in this thesis, we propose a rule-based modelling approach based on the κ platform (Danos et al., 2007) for a model of transcription. Additionally, the model offers a graph-based formalism of the transcription equipped with logical expressions that define the transitions. The logical expressions are constructed in a way to facilitate the automatic generation of executable κ rules. Finally, by simulating the rules using SSA, we obtain the mRNA copy numbers.

1.3 Contribution and organization of the thesis

Main findings of this thesis are presented in the following three chapters.

Chapter 2: Deterministic modelling approaches reveal robust behaviour of protein phosphorylation cascades

In this chapter, we employ deterministic modelling approaches based on mass-action kinetics to analyse the robustness of signalling cascades of PD cycles against external input variations. At first, we use local sensitivity and output-variance based sensitivity as measures of robustness for such cascades and observe that the efficiency of high-frequency signal attenuation increases with the number of levels in the cascades. Besides, we analyse the filtering properties of such cascades under a rigorous theoretical framework and in comparison with other PD models. Our results show that cascaded architectures behave robustly under input perturbation. In addition, cascades are able to filter out noise mimicked as high frequency signals, and thus act as a low-pass filter.

Remark. *Parts of the results (text and figures) presented in Chapter 2 are taken from the following publications by the author of this thesis:*

1. *Paul et al. (2016)*
2. *Paul and Radde (2016)*

Chapter 3: Sequestration based retroactivity as an intrinsic noise filter in protein phosphorylation cascades

This chapter carries forward our previous investigation on the robustness of protein phosphorylation cascades for the intrinsic noise due to stochasticity in molecular reactions. Using SSA, we observe that the fluctuations in terminal kinase of the PD cascade motifs are profoundly affected by cascading. Subsequently, we show that the time-varying sequestration of upstream kinase molecules is responsible for this purely stochastic effect. Besides, we determine the conditions on time scales and parameter regimes that lead to a reduction of output fluctuations. Finally, we put our results into biological context by adapting the rate parameters as well as the number of reacting molecules to a biologically feasible range for general binding-unbinding and PD mechanisms. Overall, the numerical results presented in this chapter reveal a novel role of stochastic sequestration for dynamic noise filtering in signalling cascade motifs.

Remark. *Parts of the results (text and figures) presented in Chapter 3 are taken from the following publication by the author of this thesis:*

1. *Paul and Radde (2018)*

Chapter 4: Robustness in gene expression - a rule-based approach

Chapter 4 discusses the idea of robustness in gene expression and presents the preliminary sketch of a rule-based modelling approach based on the κ platform (Danos et al., 2007) for a model of transcription in order to understand the mechanisms of robustness in terms of mRNA distributions and transcriptional bursts. The modelling approach offers a graph-based formalism, where a node represents the configuration of the promoter/enhancer along with the mRNA copy-number, named *microstate*, and an edge represents a transition between a pair of such states. A transition is allowed upon fulfilling a specific condition that is logically constructed from the status of the pair of the microstates. The formal structure is then translated to an executable form that consists of a set of κ rules. Finally, the κ rules are simulated using SSA to obtain the distribution of mature mRNA molecules and associated statistics thereof. The correctness of the rules is verified numerically using three simple instances of a gene regulation model against an alternative modelling approach constructed in parallel (Nam, 2018), and based on the solution of the CME obtained using generating functions (Mugler et al., 2009; Xu et al., 2016; Zhang et al., 2013).

Remark. *This is an ongoing project in collaboration with the Gunawardena lab at the Department of Systems Biology of the Harvard Medical School, Boston, MA, USA.*

Conclusion & outlook

We conclude the thesis by summarising the findings of all the chapters, discussing the thesis as a whole, and proposing a few future directions for this research.

Appendix

The appendix contains relevant derivations, mathematical concepts, and definitions for a better understanding of the content of this thesis.

2 Deterministic modelling approaches reveal robust behaviour of protein phosphorylation cascades

In this chapter, we compare the robustness of a cascade of two double PD cycles, a ubiquitous module that is present in multiple signalling pathways, with simpler activation motifs. We employ an ODE based modelling approach for chemical reaction kinetics and investigate robustness with respect to input perturbations, which mimic for example variations in receptor levels or amounts of other proteins upstream of the cascade. We introduce and discuss different robustness measures: a simple steady-state analysis based on local sensitivities, a variance-based approach and the ability to filter out spurious noise in a dynamic input scenario. Results illustrate that the signalling cascade of two double PD cycles can act as a reliable switch. It is sensitive in a small range about a threshold value, but extremely robust everywhere else. Furthermore, this motif acts as a low-pass filter that filters out spurious high-frequency signals.

2.1 Protein phosphorylation in cellular signalling

Protein phosphorylation (dephosphorylation), i.e. addition (removal) of a phosphate group (PO_4^{3-}) to a signalling protein, is one of the major post-translational modifications (PTMs) during the activation of signalling pathways (Ardito et al., 2017). During the phosphorylation process, the phosphate group is donated by an Adenosine triphosphate (ATP) molecule and the conversion is mediated via protein *kinase*. The reverse process or the dephosphorylation is mediated via phosphoprotein *phosphatase*. A schematic representation of the mechanism is given in Figure 2.1.

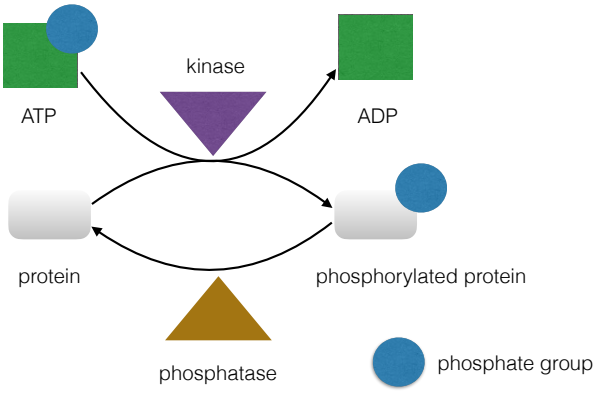


Fig. 2.1. Schematic representation of a single PD cycle. ATP acts as a *donor* of the phosphate group to the protein molecule and the phosphorylation reaction is mediated by the enzyme kinase. The reverse reaction or the dephosphorylation takes place in the presence of phosphatase molecules.

Protein phosphorylation is crucial to some important cellular processes such as signal transduction, cell growth, development and aging because activation and deactivation of many enzymes and cellular receptors are mediated via different kinases and phosphatases (Ardito et al., 2017). Abnormalities in this PTM contribute to the development of cancer cells and many other diseases (Cohen, 2001, 2002). In fact, due to this reason protein kinases and phosphatases are considered to be potential therapeutic targets for cancer treatment (Ventura and Nebreda, 2006).

Multisite protein phosphorylation

Among approximately 30% proteins that undergo phosphorylation in the eucaryotic system, many of them usually have multiple phosphorylation sites (Gunawardena, 2005). For example in the MAPK pathway as shown in Figure 1.2, MEK and ERK undergo double phosphorylation in order to transport the extracellular signal to the nucleus for gene regulation. It is well known that multisite phosphorylation significantly increases the

scope for modulating the protein function by regulating the gene expression (Whitmarsh and Davis, 2016). Multiple site phosphorylation by a single kinase regulates the binding of the transcription factor, hence affecting gene expression. Examples include activated ERK mediated phosphorylation of the ETS domain-containing protein Elk-1. When Elk-1 is phosphorylated, mediator transcription activator complex is recruited to initiate gene transcription. But, a progressive phosphorylation dissociates the mediator and recruit the repressor complex to inhibit the transcription (Mylona et al., 2016).

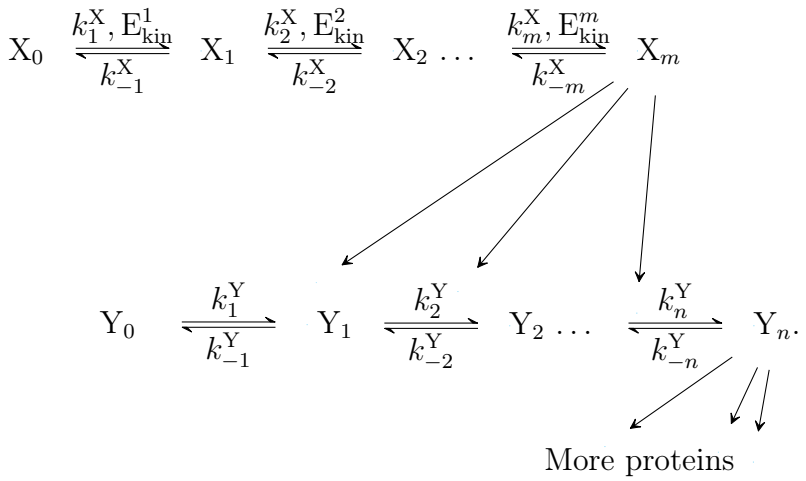
Biochemical reaction network representing multisite protein phosphorylation cascades

Consider a protein X with m phosphorylation sites, and sites are phosphorylated sequentially via m different kinases $\mathrm{E}^i_{\mathrm{kin}}$ for $i = 1 \dots m$. The reaction scheme then reads,

$$\mathrm{X}_0 \underset{k_{-1}}{\overset{k_1, \mathrm{E}^1_{\mathrm{kin}}}{\rightleftharpoons}} \mathrm{X}_1 \underset{k_{-2}}{\overset{k_2, \mathrm{E}^2_{\mathrm{kin}}}{\rightleftharpoons}} \mathrm{X}_2 \dots \underset{k_{-m}}{\overset{k_m, \mathrm{E}^m_{\mathrm{kin}}}{\rightleftharpoons}} \mathrm{X}_m.$$

Here, $\mathrm{X}_k, 1 \leq k \leq m$ is the k^{th} phosphorylated form of X, k_i and k_{-i}s for $i = 1, \dots, m$, are the deterministic reaction rate constants for forward and reverse reactions, respectively.

Consider another scenario that involves multiple proteins arranged and activated in a cascaded fashion forming a multi-tier phosphorylation cascade. For example, protein X in a fully phosphorylated form acts as a kinase for its immediate downstream protein say, Y and mediates its phosphorylation. With $k^{\mathrm{P}}_{i,-i}$ for $i = 1, \dots, m$ being the reaction rate constants for protein P, the biochemical reaction scheme for this scenario reads,



Although for both scenarios the function is *phosphorylation*, a comparatively complex architecture such as a multi-tier phosphorylation cascade involving multiple proteins is ubiquitous in major signalling pathways. For example, three-tiered cascades, which involve

three proteins that become active upon dual phosphorylation, appear in some important and well-investigated pathways such as MAPK and protein kinase B or Akt (see e.g. Brightman and Fell (2000); Fritsche-Guenther et al. (2011); O'Shaughnessy et al. (2011); Santos et al. (2007)). These pathways are characterized by various cell-type specific dynamic behaviours, including for example bistability, bimodality, graded or switch-like responses, signal amplification and ultrasensitivity (Birtwistle et al., 2012; Kholodenko, 2000; Legewie et al., 2007; Markevich et al., 2004; Qiao et al., 2007; Xiong and Ferrell, 2003). Moreover, in diseases like cancer, these pathways are found to be dis-regulated and thus serve as potential drug-targets (Grieco et al., 2013; Kolch, 2005; Kolch et al., 2005).

At this point, the question arises whether there is a mechanistic advantage, say in terms of input-output *robustness*, behind the ubiquitousness of such complex cascades. The subsequent sections are dedicated to investigate the reasons using ODE-based modelling and systems theoretic approaches.

2.2 ODE-based modelling of protein phosphorylation cascades

Assuming simple mass-action kinetics and mass conservation of total protein in the multi-tier phosphorylation cascade, the following set of differential equations describe the dynamics of protein X in the cascade,

$$\begin{aligned}
\dot{x}_1 &= k_1^{\mathrm{X}}\mathrm{E}_{\mathrm{kin}}^1\left(1-\sum_{j=1}^{m}x_j\right)+k_{-2}^{\mathrm{X}}x_2-k_{-1}^{\mathrm{X}}x_1-k_2^{\mathrm{X}}\mathrm{E}_{\mathrm{kin}}^2x_1\\
\dot{x}_i &= k_i^{\mathrm{X}}\mathrm{E}_{\mathrm{kin}}^i x_{i-1}+k_{-(i+1)}^{\mathrm{X}}x_{i+1}-k_{-i}^{\mathrm{X}}x_i-k_{i+1}^{\mathrm{X}}\mathrm{E}_{\mathrm{kin}}^{i+1}x_i \quad i=2,\ldots,m-1\\
\dot{x}_m &= k_m^{\mathrm{X}}\mathrm{E}_{\mathrm{kin}}^m x_{m-1}-k_{-m}^{\mathrm{X}}x_m,
\end{aligned} \tag{2.1}$$

where x_i denotes the concentration of the i^{th} phosphorylated form X_i normalized to the total amount of protein X. For a cascade of two doubly phosphorylated proteins i.e., $m=n=2$, we get for the dynamics of the second protein,

$$\begin{aligned}
\dot{y}_1 &= k_1^{\mathrm{Y}}x_2\left(1-\sum_{j=1}^{n}y_j\right)+k_{-2}^{\mathrm{Y}}y_2-k_{-1}^{\mathrm{Y}}y_1-k_2^{\mathrm{Y}}x_2y_1\\
\dot{y}_2 &= k_2^{\mathrm{Y}}x_2y_1-k_{-2}^{\mathrm{Y}}y_2.
\end{aligned} \tag{2.2}$$

In this context, it should be noted that Michaelis-Menten kinetics would be an alternative and more complex description for kinase and phosphatase driven reactions (Angeli et al., 2004; Gunawardena, 2005; Markevich et al., 2004). However, this is a valid approach especially for enzyme substrate reactions where the number of enzyme molecules (here the

kinases) is much smaller than the number of substrate molecules, which does not seem to be the case in the MAPK cascade, as it is argued in Gomez-Uribe et al. (2007). In this reference, a more general approach is introduced for the kinetics of a single phosphorylation event, called signalling cycle, by using a total quasi-steady-state approximation, and according to its steady state input/output behaviour, the dynamic behaviour of this cycle is classified into four different operating regimes: I) hyperbolic, II) signal transducing, III) threshold hyperbolic, and IV) ultransensitive. Our modelling approach can be seen as a simplifying linearization of regime I in this study. We note, however, that our analysis is in principle also applicable to other model versions.

Models

For our analysis, we consider the three following signalling network motifs:

Model A

$$\mathrm{X}_0 \underset{k_{-1}}{\overset{k_1,u}{\rightleftharpoons}} \mathrm{X}_1$$

Model B

$$\mathrm{X}_0 \underset{k_{-1}}{\overset{k_1,u}{\rightleftharpoons}} \mathrm{X}_1 \underset{k_{-2}}{\overset{k_2,u}{\rightleftharpoons}} \mathrm{X}_2$$

Model C

$$\begin{array}{ccccc} \mathrm{X}_0 & \underset{k^{\mathrm{X}}_{-1}}{\overset{k_1,u}{\rightleftharpoons}} & \mathrm{X}_1 & \underset{k^{\mathrm{X}}_{-2}}{\overset{k_2,u}{\rightleftharpoons}} & \mathrm{X}_2 \\ & \searrow & & \swarrow & \\ \mathrm{Y}_0 & \underset{k^{\mathrm{Y}}_{-1}}{\overset{k^{\mathrm{Y}}_1}{\rightleftharpoons}} & \mathrm{Y}_1 & \underset{k^{\mathrm{Y}}_{-2}}{\overset{k^{\mathrm{Y}}_2}{\rightleftharpoons}} & \mathrm{Y}_2 \end{array}$$

The input u, which can be time-varying or constant, describes the activity of a protein upstream mimicking an external signal. Additionally, we assume that all phosphorylations of protein X are mediated by the same kinase $\mathrm{E}_{\mathrm{kin}}$, the first protein in the cascade, whose activity is described by an input (time-varying or constant) denoted by u. The outputs of models A, B and C are denoted by $y^*_A = x_1, y^*_B = x_2$ and $y^*_C = y_2$, respectively.

The dynamics of each model is described by the following sets of ODEs:

Model A

$$\dot{x}_1 = -\left(k_1 u + k_{-1}\right) x_1 + k_1 u, \quad y_{\mathrm{A}}^* = x_1, \tag{2.3}$$

Model B

$$\dot{x}_1 = k_1 u(1 - x_1 - x_2) - k_2 u x_1 - k_{-1} x_1 + k_{-2} x_2 \tag{2.4a}$$
$$\dot{x}_2 = k_2 u x_1 - k_{-2} x_2, \quad y_{\mathrm{B}}^* = x_2. \tag{2.4b}$$

Model C

$$\dot{x}_1 = k_1 u(1 - x_1 - x_2) - k_2 u x_1 - k_{-1} x_1 + k_{-2} x_2 \tag{2.5a}$$
$$\dot{x}_2 = k_2 u x_1 - k_{-2} x_2 \tag{2.5b}$$
$$\dot{y}_1 = k_3 x_2(1 - y_1 - y_2) - k_4 x_2 y_1 - k_{-3} y_1 + k_{-4} y_2 \tag{2.5c}$$
$$\dot{y}_2 = k_4 x_2 y_1 - k_{-4} y_2, \quad y_{\mathrm{C}}^* = y_2. \tag{2.5d}$$

2.3 Analysis of robustness

In the following subsections, we investigate the characteristics of input-output robustness of the aforementioned ODE models using local sensitivity, output variance, and frequency response.

2.3.1 Local sensitivity based analysis of robustness

A vast majority of literature (Batchelor and Goulian, 2003; Blüthgen and Legewie, 2013; Caicedo-Casso et al., 2015; Dexter and Gunawardena, 2012; Dexter et al., 2015; Kirch et al., 2016; Ouldridge and ten Wolde, 2014) analyses robustness as the inverse of the amount of relative change in the steady state response upon a relative change in a constant input u. Mathematically, for a model M with steady state output $\bar{x}_M(u)$, this is captured by the normalized local sensitivity coefficient (see Appendix 6.1)

$$s_{\bar{x}}^{\mathrm{M}}(u) = \frac{\partial \ln \bar{x}_M(u)}{\partial \ln u}. \tag{2.6}$$

For the model A and B local sensitivity coefficients are derived analytically. For model A, we have

$$s^{\mathrm{A}}_{y^*}(u) = \frac{1}{1 + \frac{k_1}{k_{-1}}u}. \tag{2.7}$$

Thus, the sensitivity starts at 1 for $u = 0$ and decreases monotonically to zero with increasing input u. For model B, the steady state concentrations and sensitivities are given by

$$\bar{x}_1(u) = \frac{k_1 u}{\frac{k_1 k_2}{k_{-2}}u^2 + k_1 u + k_{-1}}, \quad s^{\mathrm{B}}_{x_1}(u) = \frac{k_1 - \frac{k_1 k_2}{k_{-2}}u^2}{\frac{k_1 k_2}{k_{-2}}u^2 + k_1 u + k_{-1}} \tag{2.8}$$

$$\bar{y}^*_{\mathrm{B}}(u) = \bar{x}_2(u) = \frac{k_2}{k_{-2}}\bar{x}_1(u)u, \quad s^{\mathrm{B}}_{y^*}(u) = \frac{\partial \ln \bar{x}_1(u)}{\partial \ln u} + 1. \tag{2.9}$$

For multisite phosphorylation of a single protein, a general expression of the local sensitivity coefficient is derived in Appendix 6.2.

For model C, the steady states are given by

$$\bar{x}_1(u) = \frac{k_1 u}{\frac{k_1 k_2}{k_{-2}}u^2 + k_1 u + k_{-1}}, \quad \bar{x}_2(u) = \frac{k_2}{k_{-2}}\bar{x}_1(u)u \tag{2.10a}$$

$$\bar{y}_1(u) = \frac{k_3 \bar{x}_2(u)}{\frac{k_3 k_4}{k_{-4}}\bar{x}_2(u)^2 + k_3 \bar{x}_2(u) + k_{-3}}, \quad \bar{y}_2(u) = \bar{y}^*_{\mathrm{C}}(u) = \frac{k_4}{k_{-4}}\bar{y}_1(u)u. \tag{2.10b}$$

For model C the sensitivity coefficients are calculated numerically. Results are shown in Figure 2.2. Steady state responses of all three models are shown on the left. The steady state output of model A increases hyperbolically, and the curve becomes sigmoidal when the cascade consists of more proteins. Remarkably, for the chosen range of u the steady state characteristic of model B lies below that of model A, which comes from the fact that in this range the intermediate single phosphorylated protein acts like a buffer.

Figure 2.3 depicts steady state responses and sensitivities for two scenarios; an increased number of tiers after addition of more PD cycles in the cascade (Figures 2.3(a) and 2.3(b)), and an increased number of phosphorylation sites for a single protein (Figures 2.3(c) and 2.3(d)). Local sensitivity analysis reveals that adding more proteins downstream of the cascade tends to be more robust (Figure 2.3(b)) , while for a particular protein, increasing the number of phosphorylation sites does not decrease sensitivity against input perturbation (Figure 2.3(d)). This observation indicates why multiple PD cycles arranged in a cascade are evolutionarily conserved signalling motifs. On the other hand, observations in Figures 2.3(c) and 2.3(d) qualitatively resemble the scenario in Gunawardena (2005, Fig.2.), which shows how ordered distributive phosphorylation and dephosphorylation becomes a good threshold but a poor switch in increasing the number of phosphorylation sites. In this

context, it should be mentioned that evolution exploits multisite phosphorylation according to the need of a specific system, and therefore it is difficult to put a threshold on the number of phosphorylation sites based on evolutionary advantage. For example, Prabakaran et al. (2012) discuss several examples of phosphorylations on more than two sites, like the cell-cycle inhibitor Sic1 for instance or the Kv2.1 potassium channel. In particular, thresholding due to multisite phosphorylation is useful for Sic1 as discussed in Klein et al. (2003). Moreover, multisite phosphorylation can be combined with other mechanisms to generate switch-like responses that are robust against stochastic or genetic variation between individuals. A scenario is discussed in Malleshaiah et al. (2010) regarding the mating decision in yeast Saccharomyces cerevisiae. The authors decipher the mechanisms through which a switch-like response is generated for shmooing[1].

[1]Shmooing is the formation of the projection through which two yeast cells join together.

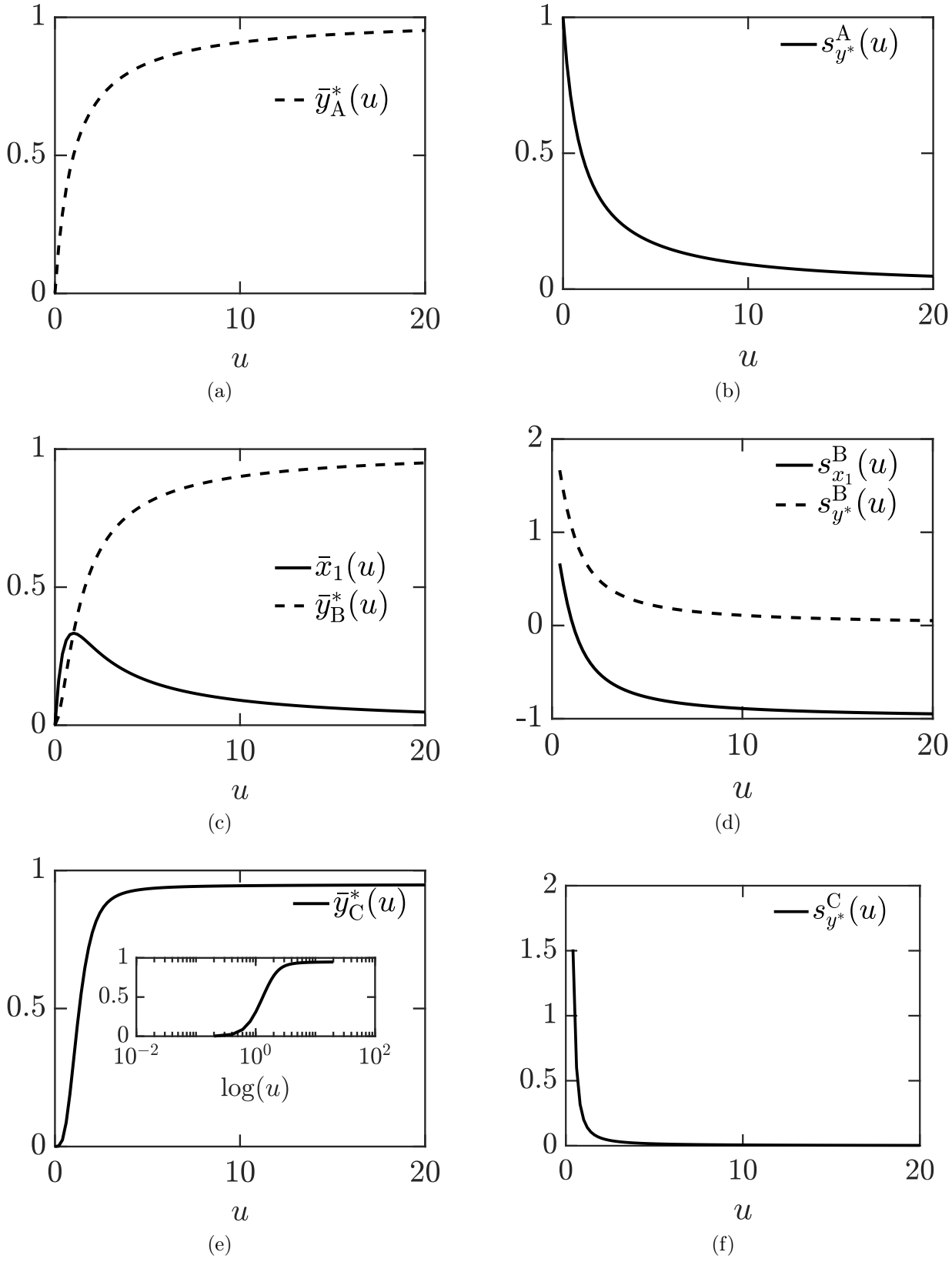


Fig. 2.2. Steady state sensitivity analysis of different model versions. (a)-(b) Steady state output characteristics and normalized local sensitivity coefficient as a function of constant input u for model A , model B (c)-(d), and for model C (e)-(f). Deterministic rate constants $k_1, k_{-1}, k_2, k_{-2}, k_{-3}, k_{-4}$, and k_3, k_4 are set to 1 and 20, respectively. These rates were chosen such that the signal is reliably propagated throughout the network.

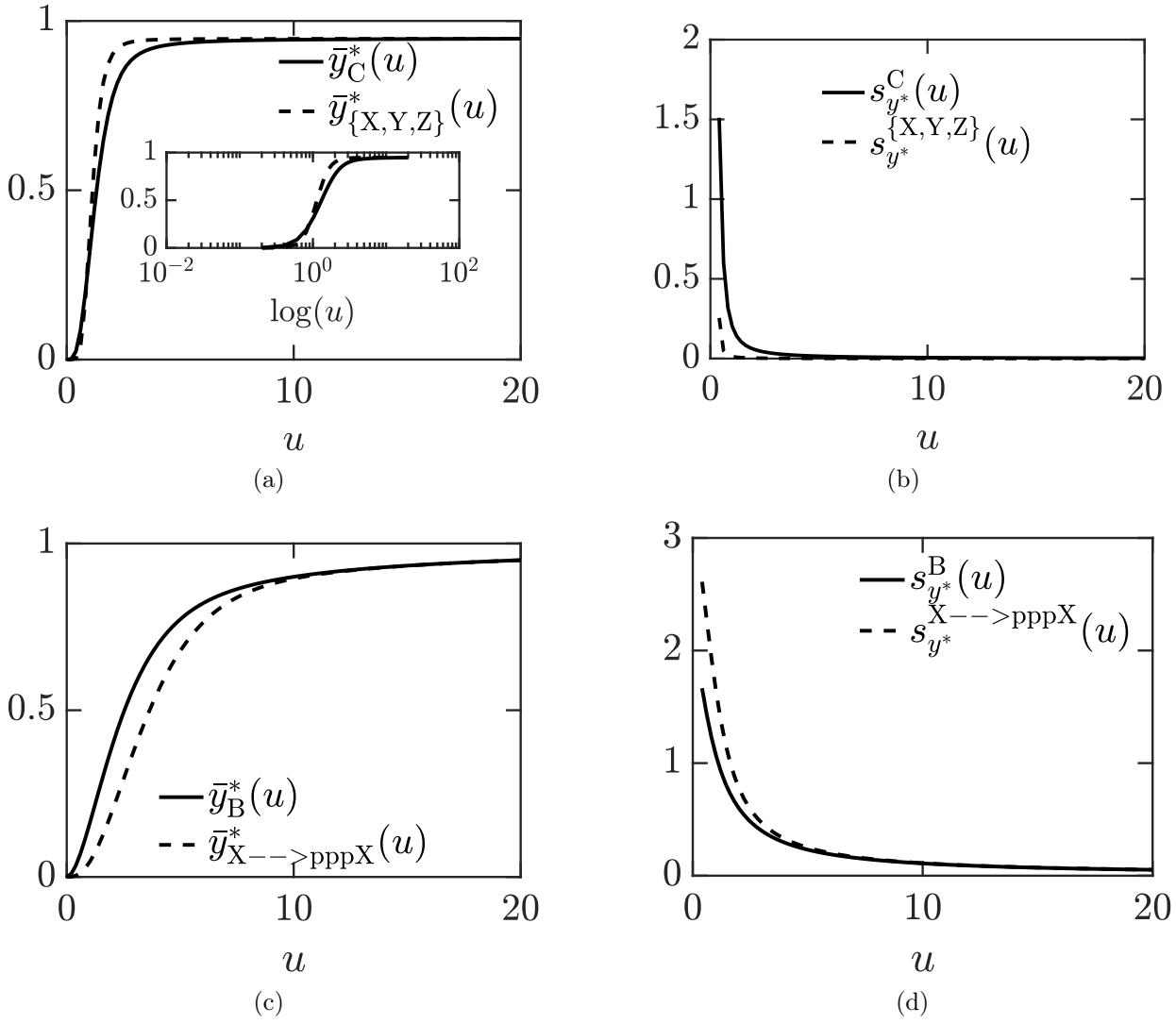


Fig. 2.3. Comparison of steady state sensitivities between more than two different proteins and a single protein with more than two phosphorylation sites. (a)-(b) model C versus cascades of three different proteins X,Y, and Z. Deterministic rate constants are set equal to that of Figure 2.2 with $k_{-5}, k_{-6} = 1, k_5, k_6 = 20$ in addition. (c)-(d) model B versus protein X with three phosphorylation sites. All the rate constants are set to 1.

2.3.2 Robustness analysis based on output-variance

As local sensitivity analysis is only suitable for small perturbations around a reference point, we complement our analysis by a variance-based approach that can take larger perturbations into account. For this purpose, the reference input variable u^r is considered to be associated with a random variable U^r having distribution $f_{U^r}(u^r)$. This approach reflects variations of the activity of the protein upstream of the cascade that might be caused by differences in receptor levels and other proteins or regulators of the signalling cascade, spatial organization or any other variability across cells. These random inputs

cause variances in model outputs, which we analyse here as functions of the reference input u^r for all three models. For illustration purposes we consider the case of a uniform input distribution, $U^r \sim [au^r, bu^r]$, with $a \in [0, 1)$ and $b > 1$, which has a coefficient of variation that is independent of the value of u^r. For model A, the distribution for $\bar{Y}_\mathrm{A}^*$ is obtained via density transformation as

$$f_{\bar{Y}_\mathrm{A}^*}^{u^r}(\bar{y}_\mathrm{A}^*) = \begin{cases} \frac{k_{-1}}{k_1(b-a)u^r(1-\bar{y}_\mathrm{A}^*)^2} & \bar{y}_\mathrm{A}^* \in [g(au^r), g(bu^r)] \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

g is a strictly monotonous function that maps reference input values u^r onto steady state outputs $\bar{y}_\mathrm{A}^*$. This distribution is illustrated in Figure 2.4(a) for different reference input values u^r. As can be seen, the support interval of $f_{\bar{Y}_A^*}^{u^r}$ is shifted to the right for increasing u^r values. Moreover, this figure suggests a bell-shaped variance as function of u^r for constant values (see Figure 2.4(c)). In fact, the respective output variance can be calculated analytically using the expression for $f_{\bar{Y}_\mathrm{A}^*}^{u^r}(\bar{y}_\mathrm{A}^*)$ in Equation (2.11).

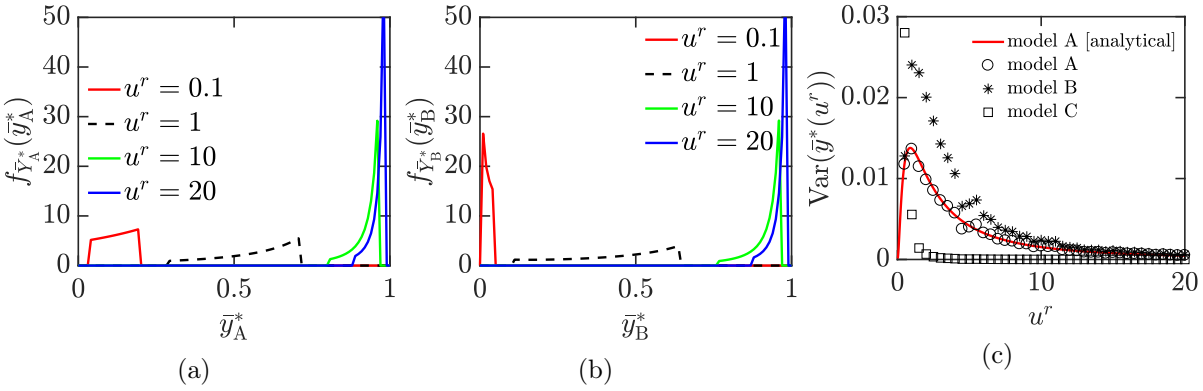


Fig. 2.4. Variance-based sensitivity analysis of models A [analytical and numerical], B, and C [numerical]. Steady state output densities for parameters $k_1 = k_2 = k_{-1} = k_{-2} = 1$ and $a = 0.4, b = 2.5$ and four different values for the reference input u^r for model A (a) and model B (b). (c) Monte Carlo approach to variance based sensitivity analysis for model A, B and C. Adapted from Paul and Radde (2016, Figs 2 and 3).

This variance-based sensitivity analysis can in principle also be applied to model B. The expression for $\bar{Y}_\mathrm{B}^*$ is (see Paul and Radde, 2016, Appendix B for a detailed derivation)

$$f_{\bar{Y}_B^*}^{u^r}(\bar{y}_B^*) = \begin{cases} \frac{1}{(b-a)u^r} \nabla_{\bar{y}_B^*} g^{-1}(\bar{y}_B^*) & \bar{y}_B^* \in [g(au^r), g(bu^r)] \\ 0 & \text{otherwise.} \end{cases} \tag{2.12}$$

The respective variance $\mathrm{Var}(\bar{Y}_\mathrm{B}^*)$ for this analysis was obtained by numerical integration. Densities for different u^r values are depicted in Figure 2.4(b). Figure 2.4(c) provides the numerically achieved variance as a function of the reference input u^r for all the three models. The variance is small for small u^r values, which is due to small input variances. In case

of model A and B, for increasing inputs u^r, the variance reaches a maximum and then decreases towards zero, because even for large input variances the output $\bar{y}_A^*$ is insensitive to perturbations in u^r for sufficiently large u^r values. The output densities and also the variance of model B have courses similar to those of model A. However, the variance of model B is about a factor $^2/_3$ higher than for model A. The variance of model C is highest for small reference inputs u^r, but then rapidly decreases to a much lower value than those of models A and B. Thus, the system response of model C is very sensitive to variations in inputs near a threshold, but becomes extremely insensitive already for input values slightly above this threshold. These results confirm that the conclusions from local sensitivity analysis are still valid for larger input variations.

In summary, local sensitivity measures together with variance-based approaches provide a simple, partly analytically verifiable and effective framework to analyse robustness of signaling cascades.

2.3.3 Analysis of robustness based on filtering properties of cascades

As a complementary approach we investigate the ability of the three models to filter out spurious noise and at the same time react reliably upon a real stimulus. Of course in reality there is no clear separation between a real signal and noisy input. However, for our analysis we assume that fast fluctuations in the input signals are more likely to mimic stochastic perturbations than slow and consistent changes. To address this question we follow the main idea described in Hersen et al. (2008), who analysed filtering properties of the HOG MAP kinase pathway experimentally and observed that for a rapidly fluctuating signal, the pathway integrates it and acts as a low pass filter, whereas for a signal varying slowly, the output of the pathway follow the input "faithfully". Therefore, we analysed the system response upon stimulation with an oscillating sinusoidal input with an offset that was chosen such that time-averaged outputs are nearly the same for all three models. After a short transient, all models show a stationary oscillating responses, which we used to estimate the output semi-amplitude (Zhou, 2013) A_i^* (half of the peak-to-peak amplitude). Results are depicted in Figure 2.5. Figures 2.5(a)–2.5(c) exemplarily show the stationary oscillations of the steady state output of model A,B, and C, respectively for two different input frequencies. For higher frequencies the low-pass filtering property of model C is clearly visible as it reduces the amplitude of the output signal significantly. In addition, Figure 2.5(d) shows a comparison of output amplitudes as functions of input frequencies for all three models. While models A and B show a very similar behaviour in this analysis, the output amplitude of model C is considerably smaller, indicating that this module responds to slow and consistent changes and ignores rapid fluctuations.

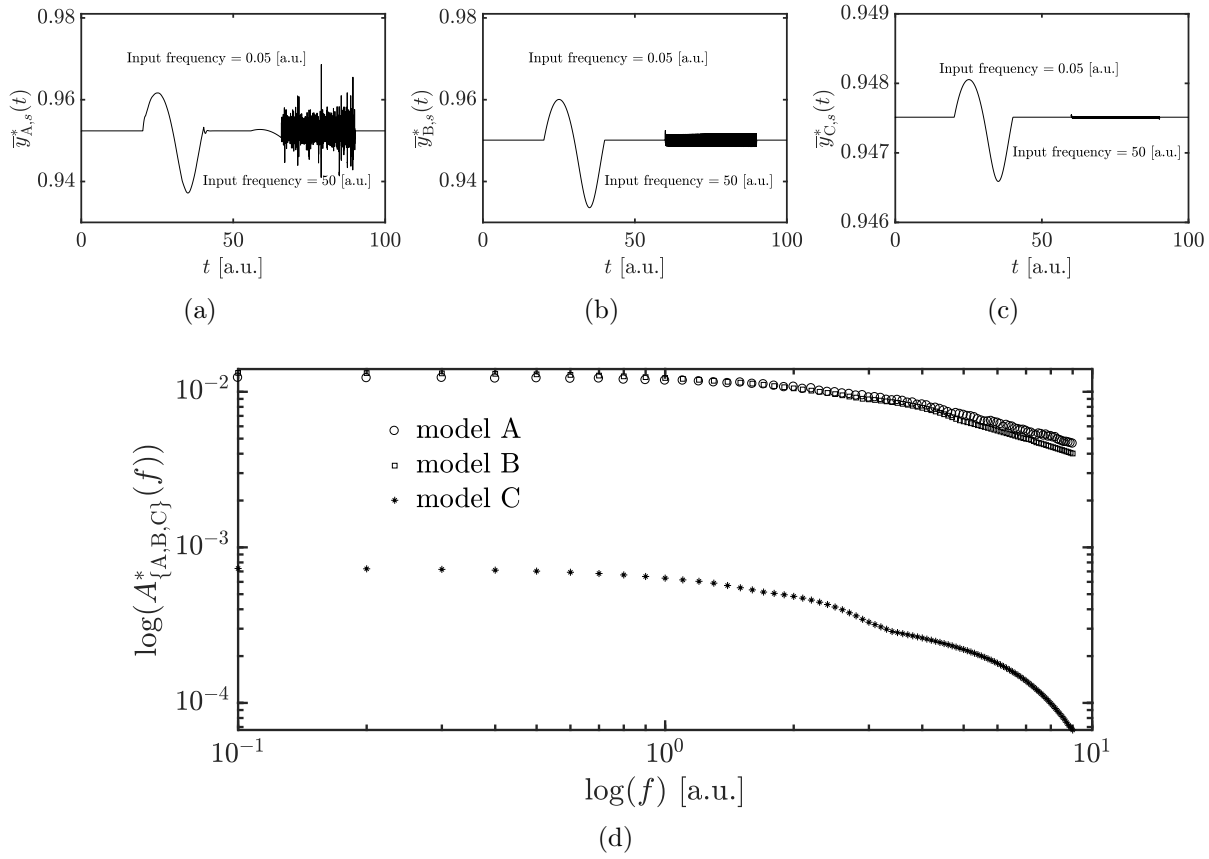


Fig. 2.5. Frequency based robustness analysis. (a)-(c) Stationary responses $\bar{y}^*_{\mathrm{A},s}(t), \bar{y}^*_{\mathrm{B},s}(t)$, and $\bar{y}^*_{\mathrm{C},s}(t)$ of model A,B, and C, respectively to oscillating input $u(t) = 20 + 5\sin(2\pi f t)$ for frequencies of 0.05 [a.u.] and 50 [a.u.]. (d) Output semi-amplitudes $A^*(f)$ for models A,B, and C. Model parameters are taken from Figure 2.2. Adopted from Paul and Radde (2016, Fig. 4)

Frequency analysis using linearization and Bode magnitude plots

In order to derive general parametric frequency analysis results, at first let us consider the following general form of a non-linear system corresponding to the models described in this chapter as

$$\dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{u})$$
$$\mathbf{y} = g(\mathbf{x}),$$

with $\mathbf{x} \in \mathbb{R}^n$ being an n dimensional state vector ($n = 1, 2, 4$ for model A, B,and C respectively), $\mathbf{u} \in \mathbb{R}^m$ is an m dimensional input vector, $\mathbf{y} \in \mathbb{R}^l$ with $1 \leq l \leq n$ being the output vector, and functions f and g are assumed to be continuous and differentiable non-linear functions. In this setting, the non-liner system described above leads to the state space representation of the following linear system when linearized around $(\bar{\mathbf{x}}, \mathbf{u}^*)$

(see Appendix 6.3 for a detailed explanation):

$$\begin{aligned} \delta\dot{\mathbf{x}} &= A\delta\mathbf{x} + B\delta\mathbf{u} \\ \delta\mathbf{y} &= C\delta\mathbf{x}, \end{aligned} \tag{2.13}$$

with $A = \nabla_{\mathbf{x}} f$, $B = \nabla_{\mathbf{u}} f$ evaluated at $(\bar{\mathbf{x}}, \mathbf{u}^*)$, $C = \nabla_{\mathbf{x}} g = (0, \ldots, 0, 1)$. The transfer function of this system is given by,

$$H(s) = C(sI - A)^{-1}B,$$

where s is the complex frequency of the form $\sigma + j\omega$ in the Laplace domain, where $\sigma, \omega \in \mathbb{R}$, $j = \sqrt{-1}$. The transfer functions for all the three models constructed in this fashion are further analysed and compared using Bode magnitude plots (see Appendix 6.4 for the definition) for two different values $u^* = 1, 10$ respectively.

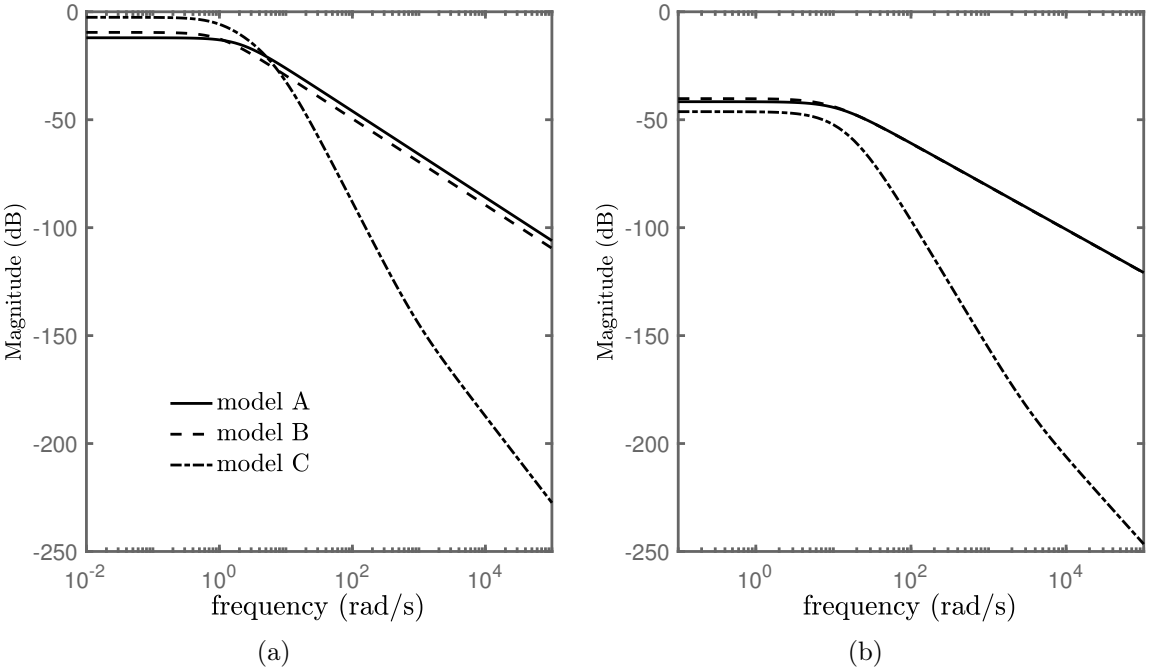


Fig. 2.6. Frequency analysis using Bode magnitude plots. (a)-(b) Bode magnitude plots for models A, B and C for input parameter $u^* = 1$ and $u^* = 10$ respectively. Original source: Paul and Radde (2016, Figs C.1 and C.2)

From the Bode magnitude plots in Figure 2.6, it is evident that all three models (A, B, and C) exhibit low-pass filtering properties because at higher frequencies, magnitudes of the corresponding transfer functions decrease. It means that the output is attenuated after a certain frequency called the *cutoff* frequency[2] for the filter. Among all the three models,

[2] a cutoff frequency is a frequency after which the output magnitude starts reducing or is attenuated. In general it refers to the frequency at which the magnitude of the output response becomes 3 dB below the value at 0 frequency

model C shows a sharper decrease in the magnitude of the transfer function after the cutoff frequency, hence proved to be more efficient in filtering out high frequency signals that are assumed to represent noise.

2.4 Summary and discussion

In this chapter we analysed robustness properties of a three-tiered phosphorylation cascade, a ubiquitous motif in intracellular signalling pathways. Moreover, we compared the robustness of its input-output behaviour with respect to various perturbations in the input signals to those of two simpler activation patterns, which are beneficial with respect to energy consumption. We explored three different concepts to investigate robustness of the model outputs against perturbations and stochastic fluctuations.

First, the systems were compared by using normalized local steady state sensitivity coefficients, a measure that describes relative changes in the output upon relative input changes. This measure is easy to evaluate and suitable for small changes around a defined reference value. It is also the most frequently used measure in other studies (Blüthgen and Legewie, 2013; Hu and Yuan, 2006; Shinar et al., 2009). Results show that the number of proteins in the cascade determine whether the response of the system is graded (which is the case for one single protein) or shows a sigmoidal behaviour (two proteins or more), with Hill coefficients that increase with increasing numbers of proteins. Ultrasensitive responses characterize efficient switches. Accordingly, sensitivities to perturbations in the input signals rapidly increase with increasing numbers of proteins in the cascade near the threshold value, while they become extremely low away from this threshold. Thus, the three-tiered phosphorylation cascade does not respond to signals that are too low, but responds reliably once the signal exceeds this threshold. While this low sensitivity far away from the threshold certainly contributes to the overall robustness, from this analysis we also expect that the system response is highly sensitive to variations in the threshold, and it remains an open question whether this is biologically desirable or balanced by other regulation mechanisms. At least, a related study in Gomez-Uribe et al. (2007) elaborates on the ability of biological systems to tune such threshold values according to the cells requirements, e.g. via gene expression regulation.

Second, we conducted a variance-based approach, in which we analysed the steady state output variance resulting from variations in the input. Unlike local sensitivity coefficients, this more global approach allows to take finite input variations into account, and thus mimics more realistic scenarios *in vivo*. Our results confirm the results from our first analysis, in which the three-tiered phosphorylation cascade was identified as an efficient switch.

Third, in a dynamic analysis approach we compared the responses of the systems to periodic inputs with different frequencies. Results show that the network motif acts as a low-pass filter that senses slow and consistent changes in the input, but filters out high frequencies. This implies that the system is able to even ignore large perturbations in form of high amplitude and high frequency input delta pulses. Although periodic inputs are probably not directly of biological relevance, we believe that such a dynamic analysis can give further insights into general response properties, going beyond steady state analysis, in particular, for transient input signals.

The simulation results presented in this study have been obtained with a predefined fixed set of parameters. These parameters were chosen such that the steady state value of the communicating intermediate (here the double phosphorylated protein) is sufficient to trigger a significant response of the following layers. In this way it is ensured that a signal is propagated reliably. Different choices of parameters might lead to signal attenuation, and hence outputs that are overall much less sensitive to input perturbations of any kind. However, taking these restrictions into account, we believe that our general conclusions are relatively robust regarding the choice of model parameters. Our approach is tightly related to the hyperbolic regime of a single signalling cycle as described in Gomez-Uribe et al. (2007), in which it is argued that at least the steady state characteristics are robust in a wide range of parameters for kinase and phosphatase concentrations (see e.g. Figure 3 in Gomez-Uribe et al. (2007)). Similarly, the output distribution of the variance-based analysis depends on the distribution of the input signal. However, the major criterion that determines the output variance is the mass of the input distribution about the threshold value in case of a sigmoidal characteristic, independent of the parametric distribution.

A next future step is to go towards more realistic modelling approaches, for example by adapting model parameters to experimental data or by embedding this motif into a larger network model, as illustrated for example in (Friedlander et al., 2015; Hu and Yuan, 2006). This direction also includes the design of suitable experiments to test our model hypothesis, which is usually a difficult step. So far, we are not aware of any existing datasets that can directly be used to validate our results. In connection with our frequency analysis, an interesting experimental device is used in (Mettetal et al., 2008), where a frequency analysis of signal transduction in the osmo-adaptation pathway is presented. Inclusion of these signalling motifs into larger networks also implies the study of feedbacks, which are omnipresent in all kinds of signalling pathways and are known to widen the range of qualitative dynamic behaviours considerably. Thus, feedback terms add another layer of complexity. However, we note here that even our simple cascade models might comprise indirect feedback effects such as competition of substrates for kinase or phosphatase molecules and sequestration effects. These could be captured by going beyond a pure mass action description (see e.g. Angeli et al. (2004)). For example, multisite-phosphorylation

can already give raise to bistability (Chickarmane et al., 2007). An interesting study in this context is also the work of Samoilov et al. (2005), who show for a simple signalling cycle module (i.e. one protein with a single phosphorylation site operating in a sigmoidal regime) that oscillations and bifurcations can already be induced by stochastic inputs, without requiring any internal feedback.

In addition to the effect of multiple proteins in the cascade, we also started to investigate the role of multiple phosphorylation sites of one protein. Their role is much less clear than that of the hierarchical cascade structure. In Gunawardena (2005) it is argued that multi-site phosphorylation creates an efficient threshold, below which the concentration of the completely phosphorylated form is nearly zero, and which increases with increasing number of sites. Moreover, to address this question, we derived analytical expressions for the local sensitivities in case of multiple phosphorylation sites of a single protein and compare them (see Figures 2.3(d) and 6.1(b)). It is evident that for smaller input, the local sensitivity increases as the number of phosphorylation sites increases and converges to the same level as that of double phosphorylation for higher input. In a different study, the double phosphorylation scheme as described in model B is known to be optimal in the sense of the approximate majority algorithm, i.e. it is the minimal motif that is required to switch a majority into a totality and decides in a fast, reliable and robust way (Cardelli et al., 2016). Moreover, since each phosphorylation consumes ATP, less phosphorylation sites are energetically favorable.

Overall, we think that the approaches presented here can in the future make valuable contributions towards a profound understanding of signalling network architecture. In particular, they could in future also be applied to investigate sensitivities towards targeted treatment options. Finally, we addressed the issue of filtering properties of a cascade in a more general way through Bode magnitude plots. In agreement with our numerical findings, model C acts as an efficient low-pass filter, as shown in the Bode magnitude plots in Figure 2.6.

3 Sequestration based retroactivity as an intrinsic noise filter in protein phosphorylation cascades

This chapter addresses the problem of robustness against *intrinsic* noise for protein phosphorylation cascades (signaling cascades mediated by protein phosphorylation). Unlike the modeling assumptions in Chapter 2, we explicitly take into account the kinase and phosphatase molecules. Our simulation results based on the SSA reveal a novel phenomenon called *dynamic sequestration* which plays an ambivalent role as an intrinsic noise filter in protein phosphorylation cascades, and has no deterministic counterpart.

3.1 Protein phosphorylation cascades and retroactivity

Recall the example of the MAPK pathway where a signal from the cell surface is transmitted to the DNA in the cell nucleus through the series of covalent modification cycles mediated by kinases and phosphatases. In recent times researchers have found that in such a signaling cascade, where one PD cycle activates the next PD cycle downstream, information does not only propagate from upstream to downstream molecules but also in the opposite direction. This effect is named *retroactivity* (Del Vecchio et al., 2008; Del Vecchio and Sontag, 2009; Ventura et al., 2009, 2010, 2008).

Retroactivity

The concept of retroactivity and retroactive effects comes from the notion of nonzero output impedance in electrical systems (Del Vecchio et al., 2008). In general, for an electrical system, impedance is a measure of the *opposition* by the system to the flow of current. In fact, in an electrical circuit, impedance is the generalization of resistance for alternating current. Therefore, impedance is frequency dependent, unlike resistance which remains constant over the frequency of the propagating current or signal. In an electrical circuit, the impedance can be of two types: input and output, depending upon whether the impedance is seen from the source or from the load perspective. Here we will consider only the output impedance as it is analogous to retroactivity. Figure 3.1 illustrates the concept of the output impedance as well as its importance in signal propagation for a cascaded system. In absence of output impedance, the voltage V' across the load element Z_L should be equal to the system's output voltage V_o. But due to output impedance, denoted by Z_o in Figures 3.1(b), V' becomes $V_o \frac{Z_L}{Z_L+Z_o}$. A formal derivation is given below.

For the circuit in Figures 3.1(b) having an alternating voltage source we can apply Ohm's law (Millikan and Bishop, 1917) as the circuit contains only resistive elements. Therefore,

$$V_o = I \cdot (Z_L + Z_o),$$

where I is the current flowing through the circuit, and

$$V' = I \cdot Z_L,$$

hence

$$\frac{V'}{V_o} = \frac{Z_L}{Z_L + Z_o} \implies V' = V_o \frac{Z_L}{Z_L + Z_o}.$$

It is clear from the expression $V' = V_o \frac{Z_L}{Z_L+Z_o}$ that when the output impedance Z_o is significantly lower than the load impedance Z_L i.e., $Z_o \ll Z_L$ then the voltage V' becomes approximately equal to the system's output voltage V_o. Hence, it is always advised to

keep the output impedance as low as possible (in fact at 0) for an electrical system. Now consider Figures 3.1(c), where two electrical systems S_1 and S_2 are arranged in a cascade. To ensure $V_o^1 = V_{\text{in}}^2$ and $V_o^2 = V_L$, Z_o^1 and Z_o^2 should be made as low as possible. Analogously, for a biological signaling cascade like MAPK (where S_1 and S_2 are single or multi-site phosphorylation cycles), it is desirable that retroactivity should be as low as possible in order to ensure unidirectional signal propagation from the upstream to the downstream modules. Retroactivity has been found to have biological context, for example in pathways

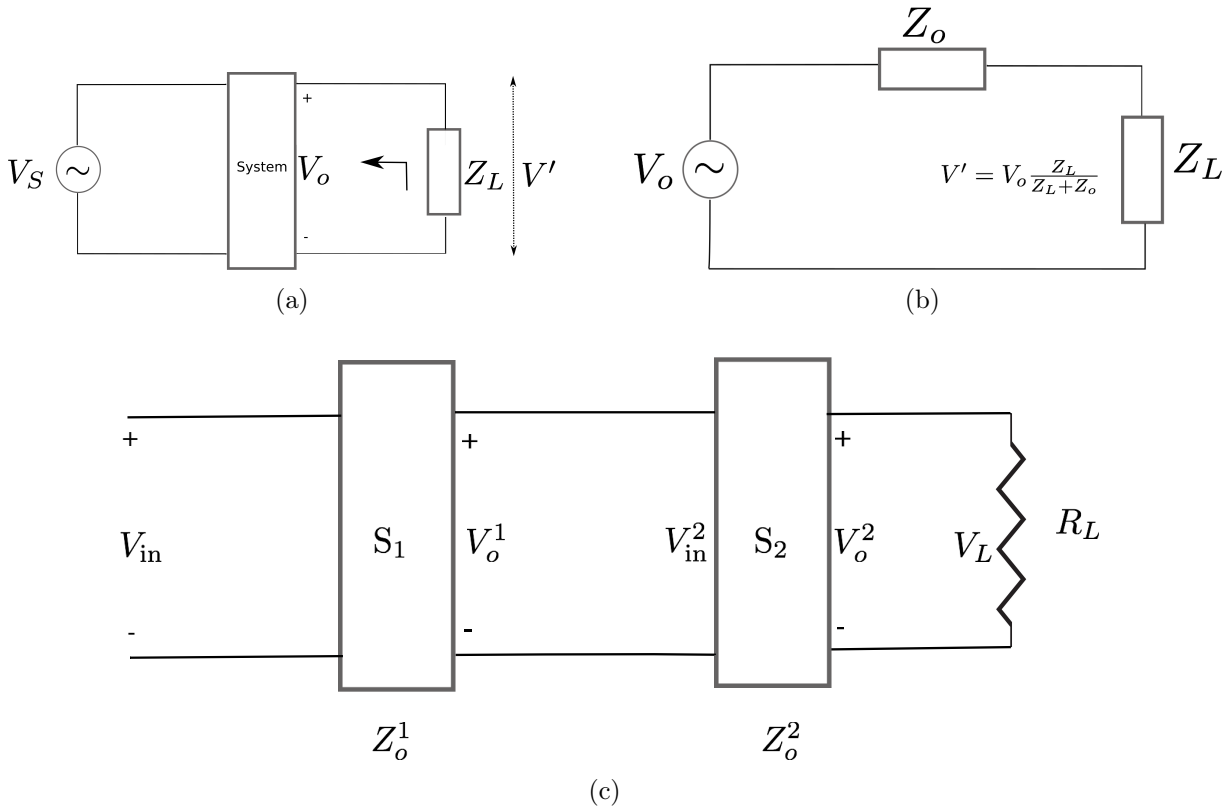


Fig. 3.1. Output impedance in electrical systems.(a)-(b) Diagrams explaining the concept of output impedance where Z_L is the load, V_S is the sinusoidal voltage source, and V' is the voltage across Z_L. (c) Two systems S_1 and S_2 are in cascades with output impedances Z_o^1 and Z_o^2 respectively. The input voltage is denoted by V_{in} and the load is denoted by the resistive element R_L.

like MAPK (Kim et al., 2010, 2011b). From a clinical perspective, retroactivity has gained much interests as it facilitates off-target effects of kinase inhibitors, a class of extremely effective anti-cancer agents (Wynn et al., 2011). Attenuation of retroactive effects is desirable (Shah and Vecchio, 2017) as it can disrupt the dynamics of the upstream modules, hence can play an important role in the dysregulation of the entire signaling pathway (Wynn et al., 2011).

On the contrary, in this work we reveal a new effect of retroactivity via sequestration of

terminal molecules which can contribute to reducing intrinsic noise and fluctuations in the activity of the terminal kinase of a double PD cycle cascade.

3.2 Models, assumptions, and simulations

For our analysis, we consider models of a simple double PD cycle motif (Figure 3.2, model A) and a cascade of two of such motifs (Figure 3.2, model B) in terms of stochastic simulations and variations in the activity of the terminal kinase. Models A and B described in Figure 3.2 are expanded using the Michaelis-Menten reaction scheme (Johnson and Goody, 2011; Michaelis and Menten, 1913),

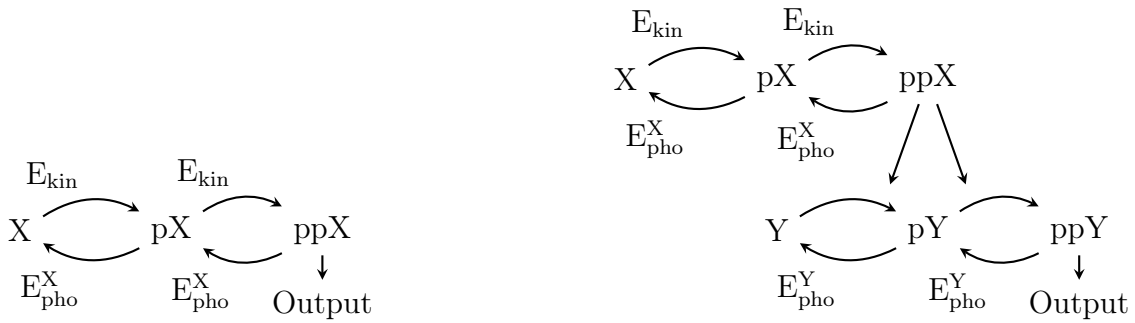
$$S + E \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftarrows}} ES \xrightarrow{k_{\text{cat}}} P + E, \tag{3.2.1}$$

where S denotes the substrate. The enzyme E is, depending on the particular reaction, either the kinase E_{kin} or the phosphatase E_{pho}. Enzyme-substrate complex and product are denoted as ES and P, respectively. Superscripts X and Y refer to the X and the Y protein modules. In model B, ppX acts as a kinase for the Y protein. The parameters k_{on}, k_{off} and k_{cat} are stochastic rate constants for binding, unbinding and catalytic reactions, respectively. Furthermore, we consider a distributive kinetics (as explained in Appendix 6.7) for double PD cycles, which requires two separate enzyme-binding events (Salazar and Höfer, 2009).

We use Gillespie's direct version of SSA (see Appendix 6.6) to generate stochastic sample paths for models A and B. Expectation values of the output and the associated variance are estimated using Monte Carlo integration of 1000 sample paths when the stochastic process is *covariance-stationary* (see Appendix 6.5 for a detailed explanation with example figures) i.e. mean and variances are time-invariant. Furthermore, we calculate CVs (see Definition 1 in Appendix 6.5) in order to quantify variability in the output. CV is a dimensionless quantity and therefore appropriate for a comparison of variables with different units or with large difference in their expectation values. For individual sample path generation, SSA is implemented in MATLAB 2016b (MATLAB, 2016). Expectation values of the output and the associated variance corresponding to 1000 sample paths are obtained using *Dizzy* (Ramsey et al., 2005).

3.3 Retroactivity via dynamic sequestration reduces output variability in cascades of PD cycles

At the beginning of our analysis we compared the CVs corresponding to the outputs of model A and B using the same set of parameters for proteins X and Y across three different



(a) model A (b) model B

Fig. 3.2. Different cascade motifs of phosphorylation-dephosphoryation cycles. (a) A double PD cycle motif (model A) and (b) a cascade of two of such motifs (model B) are compared in this study. X and Y are different proteins, pX and ppX denote single and double phosphorylated forms of protein X, and the same notation holds for the protein Y. Phosphorylation and dephosphorylation are triggered by kinase and phosphatase molecules, E_{kin} and $E^{X,Y}_{pho}$, respectively. Original source: Paul and Radde (2018, Fig. 1)

numbers of E_{kin} (Figure 3.3). As model B has more stochastic modules than model A, it is natural to expect that for model B, there will be an increase in CV because of intrinsic noise, according to Klipp and Liebermeister (2006). Interestingly, the outcome appears to be counter-intuitive in the sense that the value of c_v^{ss} for the expected output $\mathbf{E}[\text{ppY}]$ of model $B_{k^X=k^Y}$ is found to be lower than the the expected output $\mathbf{E}[\text{ppX}]$ of model A for all values of E_{kin}.

Since both modules are identical, the reduction must be caused by the different input signals those modules face. While the X protein module faces E_{kin} as a constant input, since E_{kin} is assumed a conserved quantity, the Y protein module is subject to ppX as input, which is a random variable that changes over time. The reduction in the CV could either be caused by differences in E_{kin} and $\mathbf{E}[\overline{\text{ppX}}]$, or by the fact that ppX is a time-dependent variable To determine the actual reason, at first we recorded the values of $\mathbf{E}[\overline{\text{ppX}}]$ for a range of kinase molecules, as illustrated in Figure 3.4(a). We found that for $E_{kin} = 21$ molecules both values are almost identical. Taking that into account, a comparison of the CVs of model B and of the sub-model consisting of Y protein when facing the constant input $\mathbf{E}[\overline{\text{ppX}}]$ for $E_{kin} = 21$ molecules (Figure 3.4(b) (*right*)) shows that the CV is much larger in the latter case, which supports our suggestion that the stochastic dynamics in ppX is responsible for the reduction in the CV of model B. As our primary focus is to understand the effect of sequestration dynamics on intrinsic noise, in a next step we set model B to $B_{k^Y_{\{\text{off, cat}\}}=0.01*k^X_{\{\text{off, cat}\}}}$. As a result, ppX-Y and ppX-pY complexes get accumulated due to slow unbinding and catalytic rates. In this scenario, we observe further decrease in the value of c_v^{ss}, as shown in Figure 3.5(a). This decrease in the value of c_v^{ss} confirms that the stochastic dynamic sequestration of ppX indeed controls the output variability. We repeat

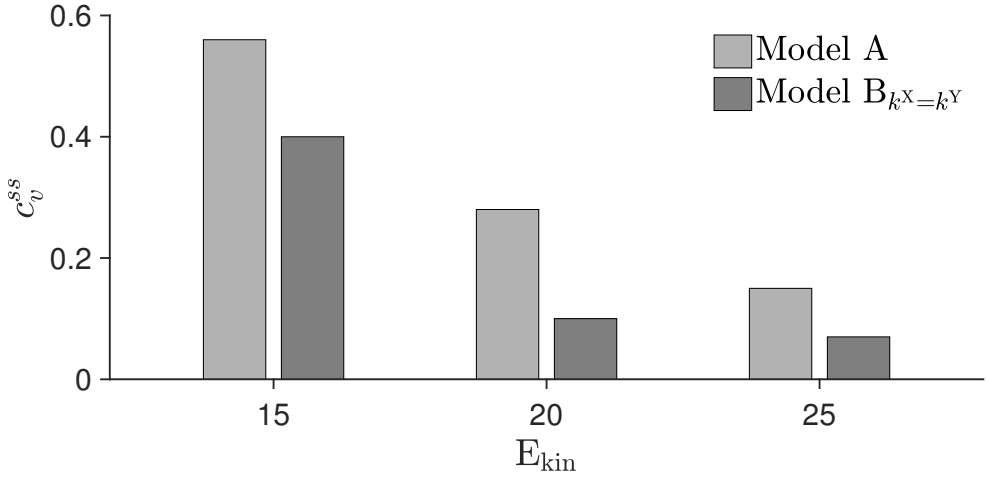


Fig. 3.3. The c_v^{ss} decreases for cascaded architectures. For models A and B, c_v^{ss} are computed for 15, 20 and 25 E_{kin} molecules from an ensemble average of 1000 SSA realizations. Parameter values are summarised in Table 3.1. Original source: Paul and Radde (2018, Fig. 2)

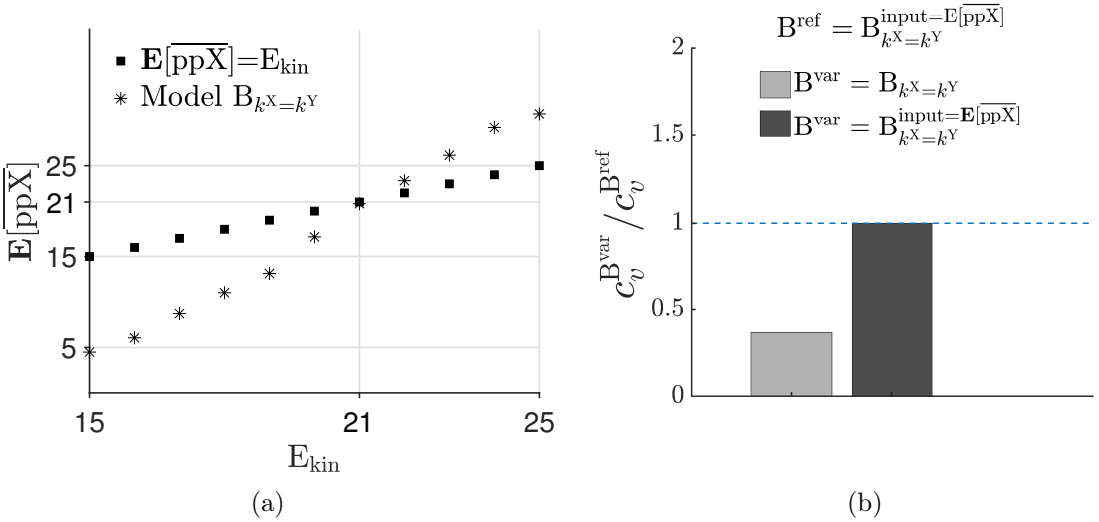


Fig. 3.4. Constant input to the Y protein module does not contribute to the reduction in the value of c_v^{ss} of $\mathbf{E}[\overline{ppY}]$). (a) $\mathbf{E}[\overline{ppX}]$ is quantified across a range of E_{kin} molecules. At $E_{kin} = 21$ molecules, the amount of free ppX takes almost the same value, providing a good reference for further comparisons. (b) Taking $E_{kin} = 21$ molecules, c_v^{ss} values are compared between model B and the case where the Y protein module of model B is fed with a constant input which is set to the expected number of free ppX molecules $\mathbf{E}[\overline{ppX}]$. The latter model variant is taken as a reference for comparison and denoted by B^{ref}. Table 3.1 summarises the rest of the parameter values. Original source: Paul and Radde (2018, Fig. 3)

the analysis shown in Figure 3.4 to explicitly exclude the differences in the mean values of the input to the Y protein module. In that way we avoid the input (to the Y protein module) to be responsible for this further reduction of the c_v^{ss} (Figures 3.5(b) and 3.5(c)).

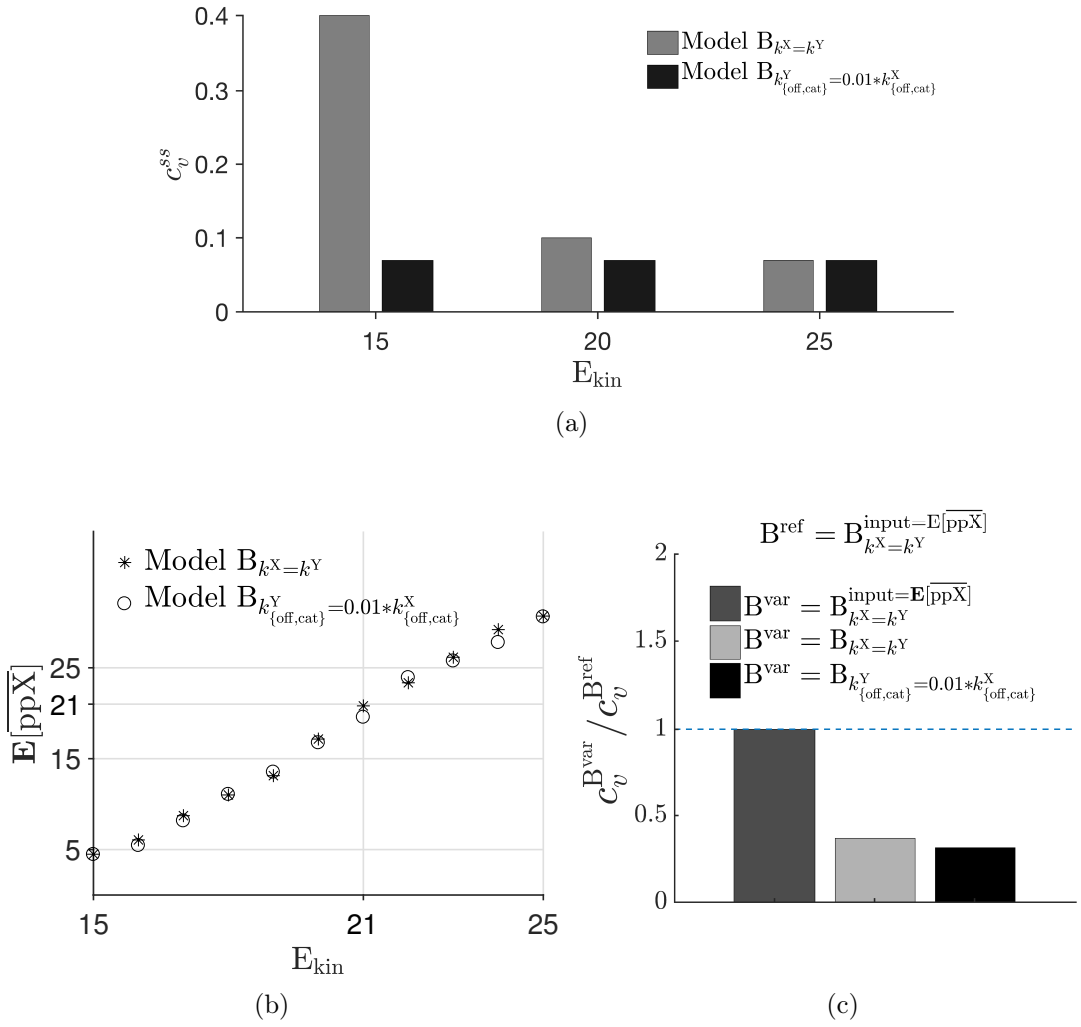


Fig. 3.5. The c_v^{ss} of the cascaded architecture further decreases for slow timescales of the unbinding and catalytic reactions (a) A comparison of c_v^{ss} for model B and model B with 100 times smaller k_{off} and k_{cat} values for the Y protein module, (b)-(c) the same analysis as in Figures 3.4 was performed to ensure that this further reduction is indeed primarily caused by the increased sequestration rate of ppX. Table 3.1 summarises the rest of the parameter values. Original source: Paul and Radde (2018, Fig. 4)

3.4 Sensing the downstream module via stochastic dynamic retroactivity

In this study, we realized retroactivity via sequestration dynamics where the sequestration of ppX molecule is regulated by the downstream molecule Y. Figure 3.6 illustrates this kind of retroactivity in the present context. Although the parameters of the X system are equal in both models, a shift in the dynamics of the X system towards ppX in the steady state is observed due to sequestration. We compare steady state expectation values of total amount

for all three X variables (X, pX, and ppX) for model A and two variants of model B (one with high and the other with a low retroactivity as described in Figure 3.6(c)). It can be seen that at steady state $\mathbf{E}[\overline{\mathrm{ppX}}^t]$ increases and $\mathbf{E}[\overline{\mathrm{pX}}^t]$ and $\mathbf{E}[\overline{\mathrm{X}}^t]$ decrease from model A to model B with low retroactivity ($\mathrm{B}^R_{\mathrm{low}}$) and to model B with high retroactivity ($\mathrm{B}^R_{\mathrm{low}}$).

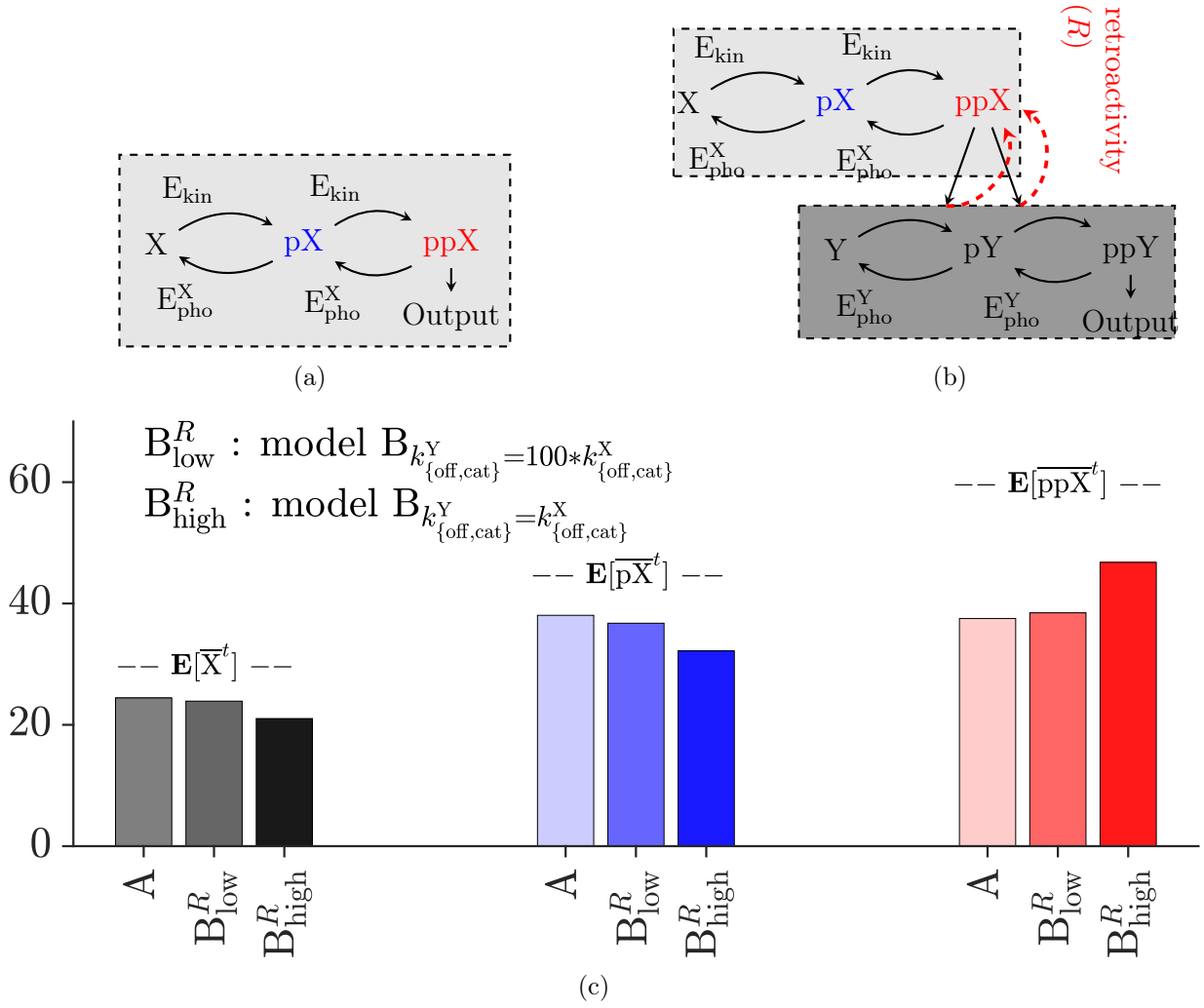


Fig. 3.6. Retroactivity via sequestration. (a) Isolated upstream module (model A). (b) Connected upstream module (model B). (c) For $\mathrm{E}_{\mathrm{kin}} = 21$ molecules, $\mathbf{E}[\overline{\mathrm{X}}^t]$ (black), $\mathbf{E}[\overline{\mathrm{pX}}^t]$ (blue), and $\mathbf{E}[\overline{\mathrm{ppX}}^t]$ (red), are estimated from an ensemble of 1000 SSA realizations, and shown for model A, model $\mathrm{B}_{k^{\mathrm{Y}}_{\{\mathrm{off,cat}\}}=100*k^{\mathrm{X}}_{\{\mathrm{off,cat}\}}}$ representing a lower retroactive effect, and model $\mathrm{B}_{k^{\mathrm{Y}}_{\{\mathrm{off,cat}\}}=k^{\mathrm{X}}_{\{\mathrm{off,cat}\}}}$ representing a higher retroactive effect. Original source: Paul and Radde (2018, Fig. 5)

Kim et al. (2011a) has provided experimental evidence of this kind of retroactivity, where the amount of doubly phosphorylated ERK in the MAPK signaling pathway has been shown to correlate with the number of ERK substrate molecules. The more substrate molecules are available, the more ppERK is sequestered by binding to these substrates.

Since those ppERK molecules are temporarily not available for the phosphatase, this sequestration affects the ratio of phosphorylated and unphosphorylated ERK towards higher phosphorylation levels. In this way, ERK can adapt its activity to the number of available substrates.

In order to investigate if a similar effect is also visible in our model setup and the X system is able to adapt to the state of the Y system, we mimicked experiments in Kim et al. (2011a) by calculating the correlation coefficient r_s between $\mathbf{E}[\overline{\mathrm{ppX}}^t]$ and the total number of the Y protein molecule or Y^T. Results are shown in Figure 3.7. From Figure 3.7, it is evident that model B is able to capture this experimentally observed behaviour, confirming that the X molecule senses the needs of the Y molecule via sequestration and adapts to the state of the Y system.

However, in our simulation scenarios in Figure 3.7, Y^T is a conserved quantity. Thus, the dynamic retroactivity cannot be explained directly by variations in Y^T. To establish the fact that the X module *dynamically* adapts the state of the Y modules via sequestration based retroactive effect, we picked up those Y molecules that are not fully phosphorylated i.e., $\mathrm{Y}^t + \mathrm{pY}^t$ - a quantity that fluctuates stochastically. We anticipate that the X protein module is able to sense these fluctuations and to adapt accordingly such that less fully phosphorylated Y molecules trigger a shift in the X protein module towards ppX, resulting in a dynamic correlation between ppX^t and $\mathrm{Y}^t + \mathrm{pY}^t$. We furthermore anticipate that the strength of these dynamic correlations is highly dependent on the dynamic range in which the whole system operates. For example, the X protein module must be fast enough compared to fluctuations in the Y protein module in order to be able to react to those changes. If this is not the case, the X protein module is too slow to adapt and fluctuations are averaged out

Subsequently we analyse the dynamic correlation between ppX^t and $\mathrm{Y}^t + \mathrm{pY}^t$ directly for the sample paths in different settings. Figure 3.8 illustrates the results. Figures 3.8(a) and 3.8(b) show representative sample paths for $\mathrm{B}_{k^{\mathrm{X}}=k^{\mathrm{Y}}}$ and $\mathrm{B}_{k^{\mathrm{Y}}_{\{\mathrm{off,\ cat}\}}=0.01*k^{\mathrm{X}}_{\{\mathrm{off,\ cat}\}}}$, respectively. Simulations were performed with a total number of $\mathrm{Y}^T = 100$ molecules. Respective distributions of correlation coefficients ρ obtained via 1000 simulation runs are shown in Figure 3.8(c). Both settings show correlations that are significantly different from zero. The scenario where timescales for X and Y protein modules are the same, we observe a negative correlation between ppX^t and $\mathrm{Y}^t + \mathrm{pY}^t$ due to the time delay in the response of the X protein module to the changes in the Y protein module. For the case where the Y protein module has a much slower dynamics, we observe a positive correlation as in that case the X protein module can follow changes in the phosphorylation state of the Y protein module instantaneously. As expected, correlations become smaller with decreasing number of Y^T molecules, as exemplarily shown in Figure 3.9, where we have used $\mathrm{Y}^T = 15$ molecules. Overall, the analysis shows that stochastic sequestration dynamics affects variability in the

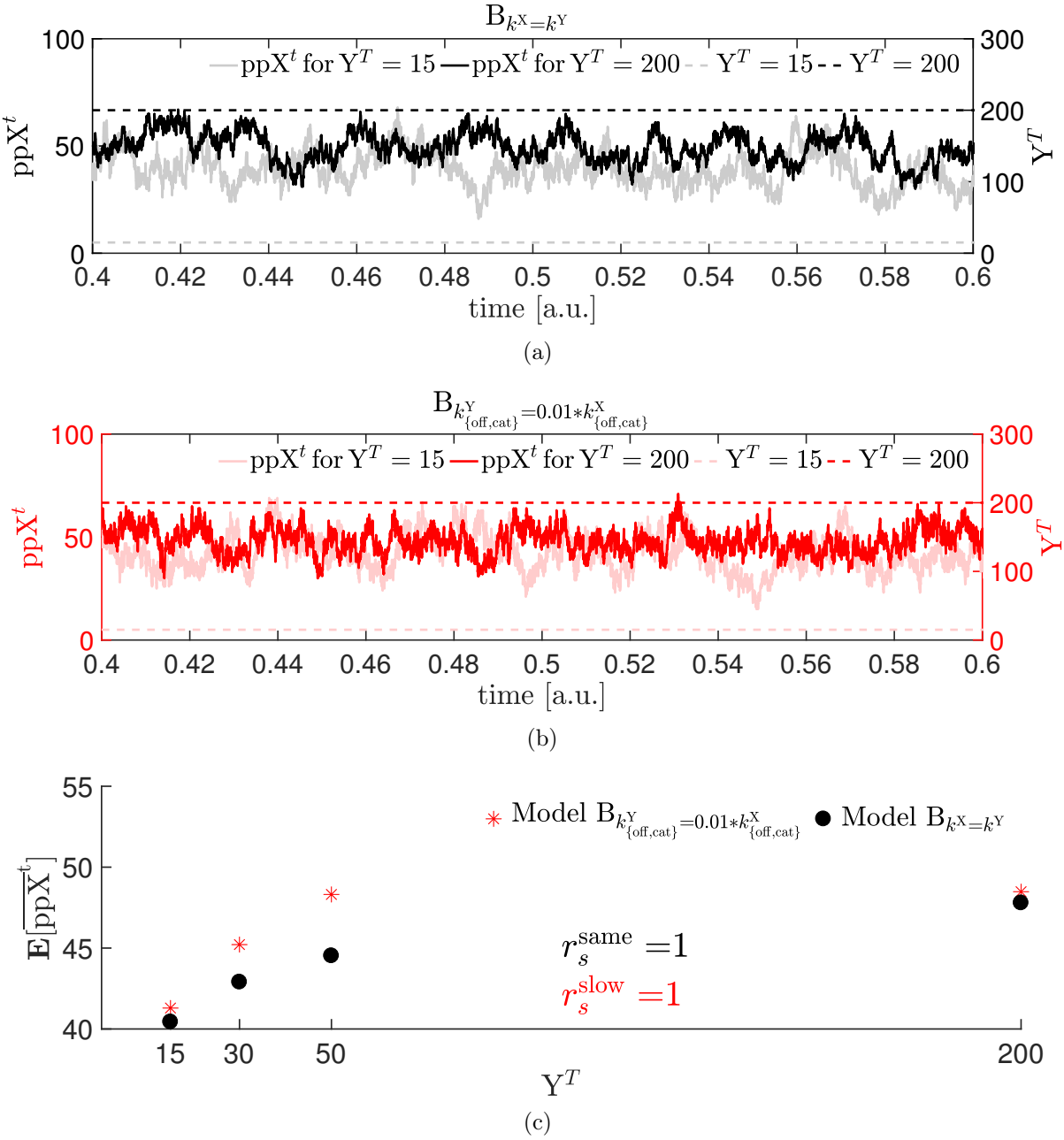


Fig. 3.7. The X protein module senses Y^T via adaptation of the sequestration rate. Representative sample paths of ppX^t for both $B_{k^X=k^Y}$ (a) and $B_{k^Y_{\{off,\ cat\}}=0.01*k^X_{\{off,\ cat\}}}$ (b) for $Y^T = 15$ (light gray and red lines, respectively) and Y^T = 200 molecules (dark gray and red lines, respectively). (c) Both settings result in a perfect rank correlation between Y^T and $\mathbf{E}[\overline{\text{ppX}}^t]$. Original source: Paul and Radde (2018, Fig. 6)

activity state of the downstream protein of cascades of double PD cycle motifs. Of note is that this form of retroactivity can only be observed in a stochastic environment and has no counterpart in a deterministic regime.

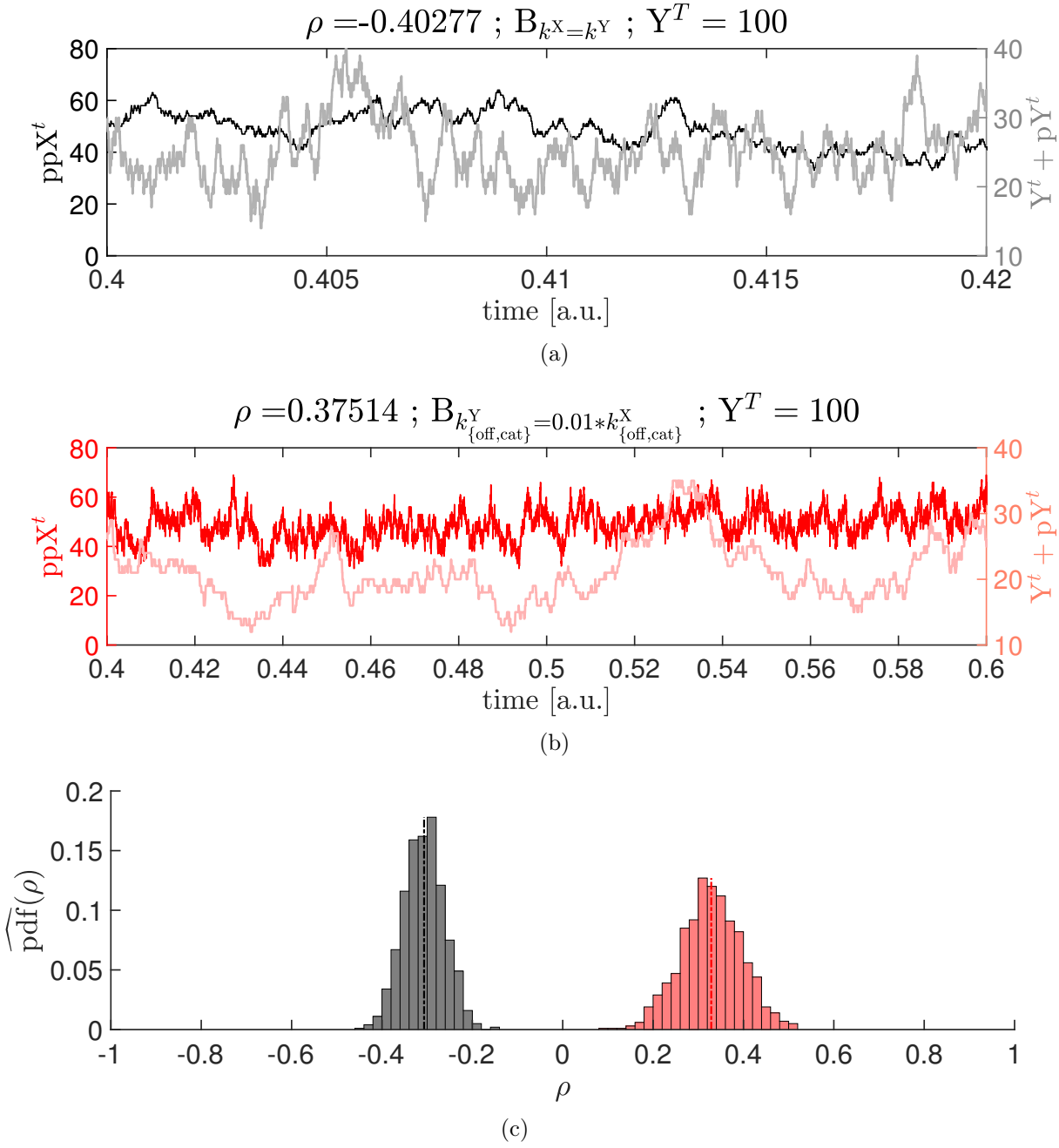


Fig. 3.8. Stochastic sequestration dynamics causes a reduction in output variability of model B. Sample paths of ppX^t and $\mathrm{Y}^t + \mathrm{pY}^t$ for $\mathrm{B}_{k^{\mathrm{X}}=k^{\mathrm{Y}}}$ (a) and $\mathrm{B}_{k^{\mathrm{Y}}_{\{\mathrm{off,\ cat}\}}=0.01*k^{\mathrm{X}}_{\{\mathrm{off,\ cat}\}}}$ for $\mathrm{Y}^T = 100$ molecules (b). (c) Distributions of correlation coefficients ρ for both settings that have been inferred via 1000 SSA simulations. Parameter settings are listed in Table 3.1. Parameter values: $\mathrm{Y}^T = 100$, $\mathrm{E}_{\mathrm{kin}} = 21$ molecules, $\mu_1 = -0.3068$, $\sigma_1 = 0.0447$ (black) $\mu_2 = 0.3287$, $\sigma_2 = 0.0672$ (red). μ_i and σ_i for $i = \{1, 2\}$, denote the mean and variance of distributions of correlation coefficients ρ, respectively. Original source: Paul and Radde (2018, Fig. 7)

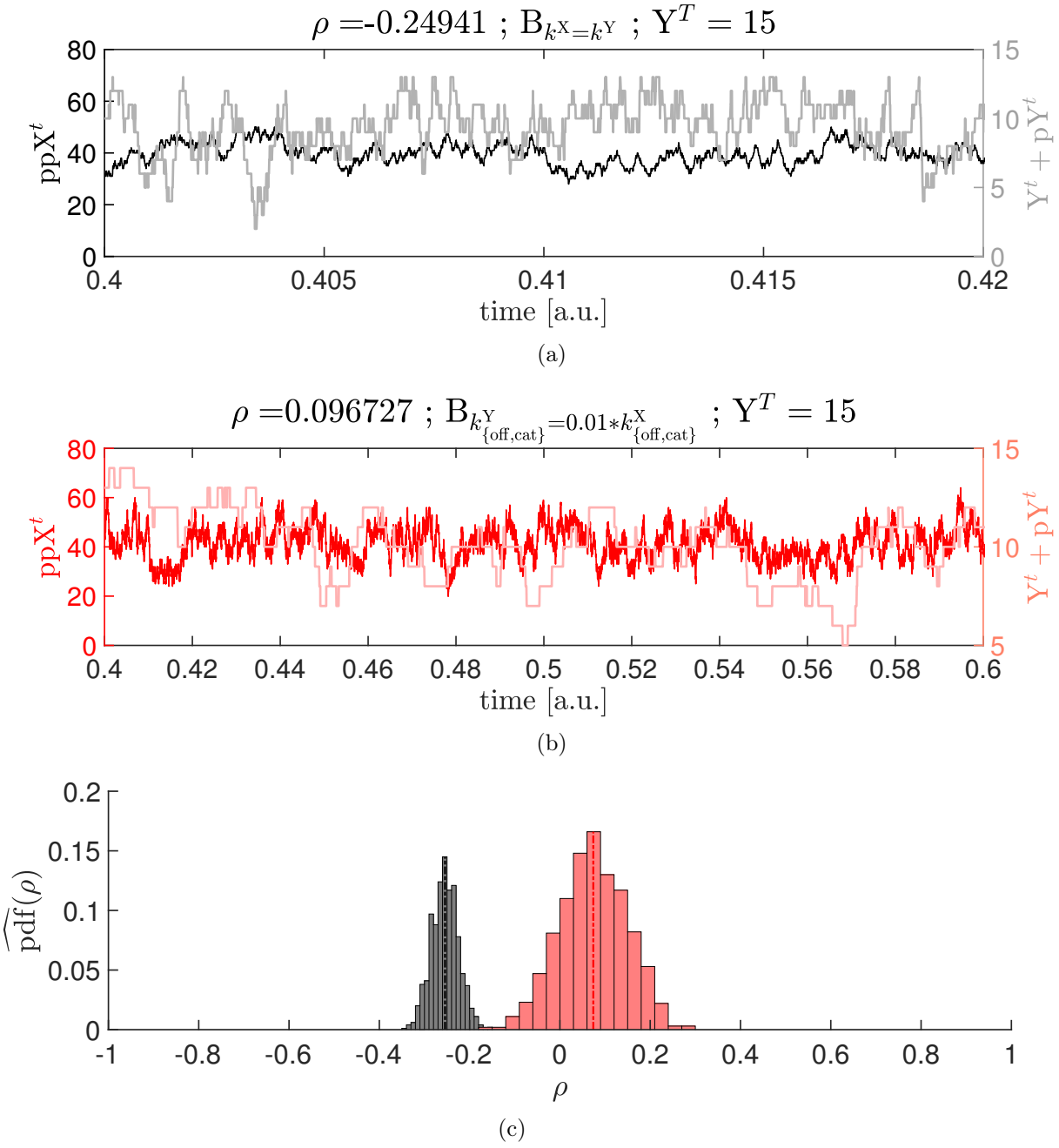


Fig. 3.9. Parameter values: $Y^T = 15$, $E_{kin} = 21$ molecules, $\mu_1 = -0.2544$, $\sigma_1 = 0.0297$ (black), $\mu_2 = 0.0739$, $\sigma_2 = 0.0745$ (red). Original source: Paul and Radde (2018, Fig. A.10)

3.5 Dynamic sequestration in biological systems

In previous sections, we observed how sequestration dynamics is attenuating intrinsic noise for signaling cascades. However, the observation is restricted to the chosen parameter set which may not be relevant in a biological context. Therefore, we decide to set model parameters within biologically feasible ranges where applicable. For this purpose, we adopt our parameters by using values recorded in Table 1 of Dhananjaneyulu et al. (2012). These

Initial number of molecules	X_{init}	Y_{init}	E^X_{pho}	E^Y_{pho}
	100	100	20	20
Stochastic rate constants	k^X_{on}	k^X_{off}	k^X_{cat}	
	0.01	0.02	0.08	

Table 3.1. Parameters for Figures 3.3–3.5, 3.7–3.9, 3.11(a), 3.11(c) and 3.11(e). Stochastic rate constants k^X_{on} (binding), k^X_{off} (unbinding) and k^X_{cat} (catalytic) have units of (molecule^{-1}time^{-1}), (time^{-1}) and (time^{-1}) respectively. The superscripts denote the respective protein module. For example, k^X_{on} denote the binding rate for protein module X. Original source: Paul and Radde (2018, Table 2)

values have been inferred from experiments on the Ras/MEK/ERK signaling cascade in human HeLa cells, as described in Fujioka et al. (2006). Resulting parameter values are listed in Table 3.2. Using these values, we performed the same analysis as in Figure 3.3. Results are recorded in Figure 3.10. While Figure 3.10(a) clearly shows the reduction in c_v^{ss} from model A (denoted as A^{bio}) to model B (denoted as B^{bio}) and model B with reduced rate constants for the Y protein module ($B^{bio}_{k^Y_{\{off,\ cat\}}=0.01*k^X_{\{off,\ cat\}}}$) for the range of $E_{kin} = 60 - 90$ kinase molecules, model B has a considerably higher output variability for the case $E_{kin} = 50$ molecules as compared to model A. The superscript "bio" indicates the biological context. We explain this behaviour using sensitivity analysis via dose response curves, in which we analyse model outputs with respect to different values of E_{kin}. Figure 3.11 illustrates the results for 'non-biological' context having random parameters (left column) and the 'biological context' (right column) where parameters are adopted within a biologically feasible range. The left column shows that $\mathbf{E}[\overline{ppX}]$ and $\mathbf{E}[\overline{ppY}]$ are both in a highly dynamic range for E_{kin} from 15 to 25 molecules (indicated by shaded regions). For this regime, $\mathbf{E}[\overline{ppX}]$ approximately spans a range between 4 and 44 molecules, the respective range for $\mathbf{E}[\overline{ppY}]$ is between 30 and 60 molecules. Thus, both variables are highly sensitive to variations in the input, although $\mathbf{E}[\overline{ppY}]$ to a lesser extent. For $B_{k^Y_{\{off,\ cat\}}=0.01*k^X_{\{off,\ cat\}}}$, the $\mathbf{E}[\overline{ppY}]$ curve increases much faster and is in saturation at about $\mathbf{E}[\overline{ppY}] = 60$ molecules already at $E_{kin} = 10$.

Thus, $\mathbf{E}[\overline{ppY}]$ is extremely insensitive to variations in E_{kin} and to stochastic fluctuations in ppX. The results for simulations in the biological context are illustrated on the right. Especially for $E_{kin} = 50$ molecules, $\mathbf{E}[\overline{ppX}]$ is still at an extremely low value with a small sensitivity, while $\mathbf{E}[\overline{ppY}]$ has just reached the start of its dynamic range and thus shows a high sensitivity with respect to variations in E_{kin}, which explains the increase of the CV in Figure 3.10(a). As before, for $B^{bio}_{k^Y_{\{off,\ cat\}}=0.01*k^X_{\{off,\ cat\}}}$, $\mathbf{E}[\overline{ppY}]$ has reached saturation for the range of E_{kin} values that are considered here and hence shows extremely low sensitivities and low coefficients of variation. For higher E_{kin} values, $\mathbf{E}[\overline{ppX}]$ rapidly comes into its dynamic range, while $\mathbf{E}[\overline{ppY}]$ is already almost saturated for $E_{kin} = 60$ molecules, explaining the

Initial number of molecules	X_{init}	Y_{init}	E^X_{pho}	E^Y_{pho}
	757	567	32	32
Stochastic rate constants	k^X_{on}	k^X_{off}	k^Y_{on}	k^Y_{off}
phosphorylation	0.0016	0.01	0.0021	0.01
dephosphorylation	0.0141	0.01	0.0141	0.01

Table 3.2. Parameters for Figures 3.10, 3.11(b), 3.11(d) and 3.11(f). In Dhananjaneyulu et al. (2012), the values of the Michaelis-Menten (MM) constants of phosphorylation and dephosphorylation reactions for both the X and the Y potein modules are given together with the catalytic rate constants. k^X_{on} (binding), k^X_{off} (unbinding) and k^X_{cat} (catalytic) have units of (molecule^{-1}time^{-1}), (time^{-1}) and (time^{-1}) respectively. Here, first a value of $k_{off} = 0.01$ time^{-1} is taken, which is within the range $[10^{-3}, 10^{-1}]$ for a typical mammalian cell (Milo, 2013). Subsequently, respective values for k_{on} are calculated using the relation $K = \frac{k_{off}+k_{cat}}{k_{on}}$, where K is the MM constant. For the X system, we denote the MM constants for phosphorylation and dephosphorylation reactions by K^X_{pho} and K^X_{depho}, respectively. The same notation applies for the Y system. The values for $K^X_{pho}, K^X_{depho}, K^Y_{pho}$, and K^Y_{depho} were set to $120, 22, 110$, and 22 molecules, respectively, according to Dhananjaneyulu et al. (2012). The corresponding k_{cat} values are $0.18\,s^{-1}, 0.3\,s^{-1}, 0.22s^{-1}$ and $0.3\,s^{-1}$, respectively (Dhananjaneyulu et al., 2012). Values for the range of the number of kinase molecules chosen here, $E_{kin} = [50, 60, 70, 80, 90]$ is in the same order of magnitude as the value $E_{kin} = 94$ recorded in Dhananjaneyulu et al. (2012). Original source: Paul and Radde (2018, Table 3)

immense decrease in the CV in model B^{bio} from $E_{kin} = 50$ to $E_{kin} = 60$ molecules. Taken together, this analysis shows that the dynamic range in which the system operates is crucial for the effect of cascading on the output variation and also highly influences stochastic sequestration dynamics.

3.6 Summary and discussion

In this chapter we compared a double PD cycle model (model A) and a cascade of two of such models (model B) with respect to stochastic variations in the activity of the downstream protein. Our analysis revealed an ambivalent role of stochastic sequestration dynamics for the regulation of the Y protein module variability, here measured in terms of coefficients of variation in doubly phosphorylated Y, ppY. Sequestration of doubly phosphorylated X, ppX, by the Y protein module constitutes a kind of retroactivity. Via sequestration, the X protein module senses and reacts to the state of the Y protein module, and hence information is propagated from the downstream to the upstream molecules in these protein cascades. This effect causes a correlation in the sample paths of ppX^t and those molecules of the Y system that are not fully phosphorylated, Y^t and pY^t, and results in a reduction of the CV of ppY in most of the cases that we considered. Moreover, we also investigated conditions for stochastic dynamic sequestration to have a notable effect, which highly

depends on the dynamic range in which the whole system operates. We argued that the time scale of the X protein module must be fast enough such that it can dynamically adapt to changes in the state of the Y protein module, otherwise those changes are averaged out and the correlation in the sample paths disappears. Moreover, the sequestration rate of ppX must have an impact on the X protein module, which is for example not the case if the total number of Y molecules, Y^T, is too small. Depending on operating regimes in the dose response curves of the system, we revealed that dynamic sequestration can also have the opposite effect, namely enhancing output variability. This is the case if the system operates in a regime where ppY is highly sensitive to changes in E_{kin} and at the same time the X protein module is too slow to react instantaneously to state changes in the Y protein module. In this case we observed stochastic oscillations (results not shown) in ppY around its nominal value. The Y protein module reacts sensitively to stochastic changes in ppX, and the response of the X protein module lacks behind. Similar to a negative feedback with a time delay, this leads to oscillating behaviour in the state of the Y protein module, and the variation in ppY is increased in this particular case.

So far, retroactive effects have mainly be studied via deterministic approaches. In a recent study (Shah and Vecchio, 2017) it was mathematically shown that retroactivity is attenuated in cascaded phosphorylation and phosphotransfer systems with single and/or double PD cycles with kinase as input, when maintaining a low-high substrate concentration pattern like the MAPK signaling model in Huang and Ferrell (1996). The same architecture with substrate as input is incapable of attenuating retroactivity. Until now, different effects of retroactivity have been described, including the conversion of a graded response into a switch-like response in the context of transcription factor decoy sites (Lee and Maheshri, 2012). Of note, retroactivity via stochastic dynamic sequestration has no direct deterministic counterpart, and it remains a challenging question for the future whether its effect is relevant in real biological systems.

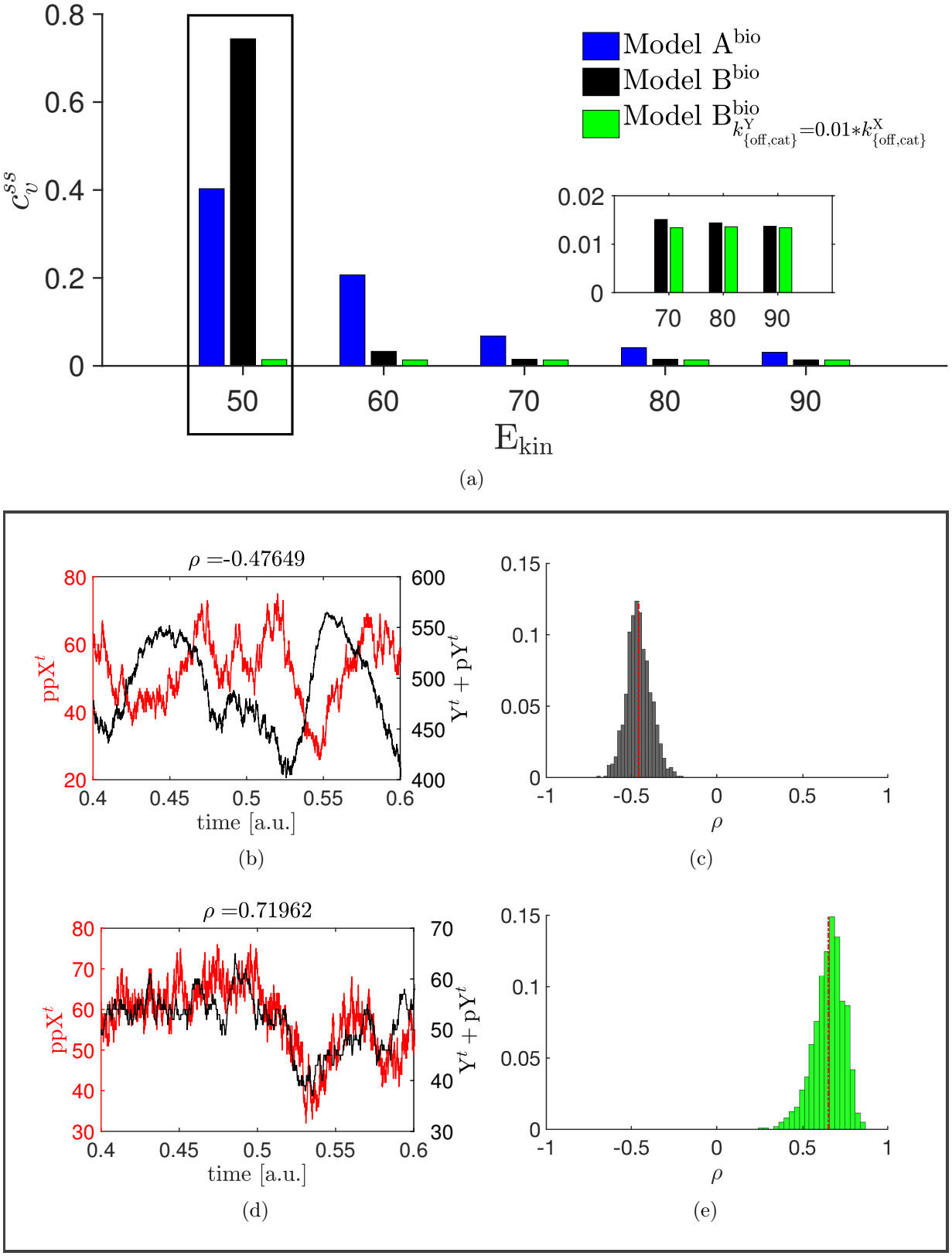


Fig. 3.10. Stochastic sequestration dynamics in a biological context. (a) Same analysis as in Figures 3.3. Different model variants are compared in terms of their coefficients of variations c_v^{ss}. (b)-(e) Sample path and correlation analysis for $E_{kin} = 50$ molecules for model B^{bio} (b-c) and model $B^{bio}_{k^Y_{\{off,\ cat\}}=0.01*k^X_{\{off,\ cat\}}}$ (d-e). Parameters are recorded in Table 3.2. Original source: Paul and Radde (2018, Fig. 8)

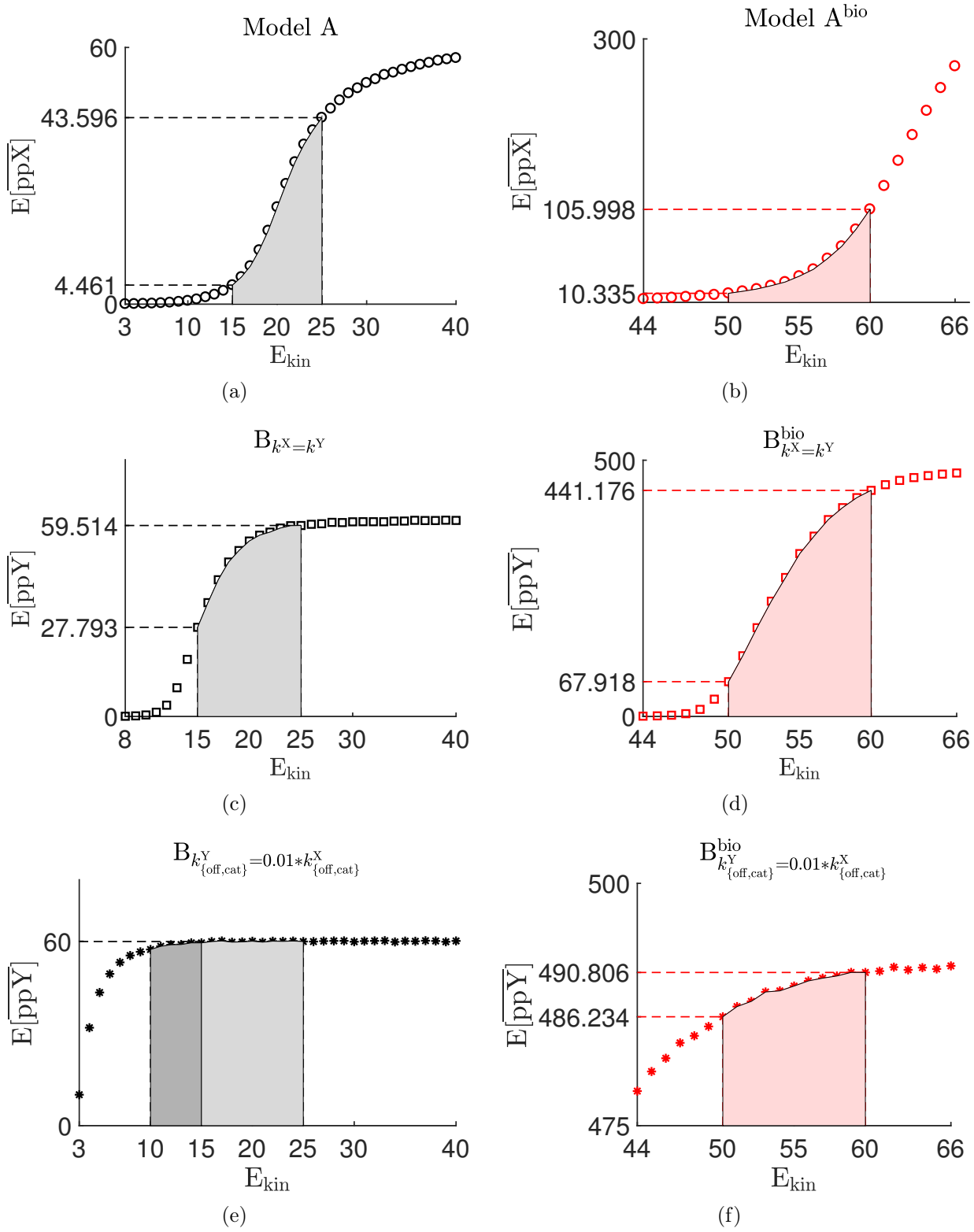


Fig. 3.11. Dose-response curves for model outputs. Expectation values for model outputs as functions of E_{kin} for the 'non-biological' context (left), and the biological context (right). Original source: Paul and Radde (2018, Fig. 9)

4 Robustness in gene expression - a rule-based approach

The source of phenotypic variations among individuals in a population is the stochasticity in gene expression that is manifested as mRNA and protein bursts at the level of transcription and translation, respectively (Kumar et al., 2015). This phenotypic variation is necessary for evolvability, which is the ability to introduce novel adaptations (Payne and Wagner, 2015).

In this respect, the goal of robustness is the opposite to that of stochasticity as the former preserves the adaptive traits in the presence of genetic and environmental changes. We have discussed this paradox in the introduction and understood that robustness favours variability at the individual level, and at the same time variability increases robustness at the population level. In that sense, stochasticity and robustness are not mutually exclusive (MacNeil and Walhout, 2011). There are multiple pieces of evidence where stochasticity favours phenotypic robustness. For example, as reviewed in Roignant and Treisman (2009), during the eye development in *Drosophila*, selection of a cell in each ommatidium (optical unit for compound eye in insects) for R8 photoreceptor prevents other cells from becoming R8 photoreceptors and directs the assembly of other photoreceptor cells in each ommatidium. Stochasticity guides the selection of a cell by providing a wide range of expression levels for a group of genes to cross a threshold level required for differentiation of that particular cell. Therefore, in this example, stochasticity is used to generate a robust phenotype. At the level of transcription, it has been observed that preloaded RNAPII at the promoter region can minimise stochasticity across a tissue to facilitate proper developmental timing and coordination by accelerating induction of gene expression (Boettiger and Levine, 2009; MacNeil and Walhout, 2011). It is evident from the above discussion that in order to understand the mechanisms of robustness in gene expression, in particular at the level of transcription, we need to understand the mechanisms of stochasticity in terms of mRNA copy numbers, bursting parameters (burst size and frequency)[1]. By understanding the mechanisms of stochasticity in terms of mRNA copy numbers and bursting parameters, we mean how these quantities are modulated by mechanisms that are associated with

[1]Burst frequency is the number of bursts per time units, and burst size is the mean number of transcripts produced per burst episode (Nicolas et al., 2018).

transcription, such as chromatin looping, TF binding, promoter-proximal pausing, RNAPII elongation dynamics. In the previous two chapters (Chapters 2 and 3), using deterministic (based on ODEs) and stochastic modelling (based on Monte-Carlo simulations) approaches, we demonstrated mechanisms by which protein signalling cascades ensure robustness against input perturbations. In this chapter, we take a digression from the ODE-based approach and present a rule-based modelling approach based on the κ platform with the primary aim to identify patterns in mRNA copy number distributions at steady state and bursting parameters across the set of regulatory mechanisms mentioned above.

4.1 Transcriptional bursts - the source of stochasticity in gene expression

Gene expression at the level of transcription occurs in pulsatile bursts as transcription switches between "on" and "off" states (Golding et al., 2005). Such bursting nature of transcription is considered as the primary source of transcriptional noise - the cause of variability in a population. Therefore, regulation of bursting parameters is necessary to control the noise or phenotypic variation in a population. For example, in a single cell experiment with *Bacillus subtilis*, the authors in Ozbudak et al. (2002) observed that for a gene with low transcription rate (the rate at which mRNA transcripts are produced) and high translation rate (the rate at which mRNA molecules are translated to protein molecules), produces large, variable and infrequent bursts, resulting in a higher phenotypic variation in the population. Conversely, with high transcription rate and low translation rate, bursts are smaller in size and frequent as well, resulting in smaller phenotypic variation in the population. For example, Senecal and co-authors (Senecal et al., 2014) investigated the impact of TFs in the bursting behaviour on a long timescale. The authors found out that for the *c-Fos* gene (a proto-oncogene i.e. a normal gene having the potential to become an oncogene that causes cancer, upon mutations and increased expression) during MAPK induction, TF concentration modulates the burst frequency, but the burst size remains unchanged. Hence, in this particular example, robustness is achieved concerning the burst size against perturbations in the TF concentration. In another work, Bartman and co-authors (Bartman et al., 2016) carried out single-molecule fluorescence *in-situ* hybridization (smFISH) experiments[2] for the β-globin gene[3] to observe the effect of chromatin looping via β-globin enhancer in modulating the bursting parameters. The authors found out that increasing the frequency of looping increases the burst fractionbut not the burst size. In an earlier study, Raj and co-workers (Raj et al., 2006) carried out a smFISH experiment with Chinese hamster ovary cells, where they reported an increase in the burst size with increasing levels of TFs, while the frequency remained unchanged. From all the examples mentioned above, it is clear that the bursting kinetics varies widely across eukaryotic cells (Nicolas et al., 2017; Suter et al., 2011). In this context, Nicolas et al. (2017) provides a non-exhaustive yet highly informative list of literature concerning the modulation of bursting parameters under different molecular mechanisms associated with transcription such as chromatin looping, availability of TFs, histone modifications, nucleosome occupancy,

[2]Instead of providing a an average measure of transcripts across a population of cells, smFISH provides information about transcripts localized in cells, so that one can analyse several different transcripts simultaneously (Kwon, 2013).

[3]β-globin gene codes for β-protein. The β-protein is a part of the larger protein Haemoglobin that is responsible for oxygen transport in red blood cells of almost all vertebrates.

number of *cis*-regulatory elements (regions of non-coding DNA found in the vicinity of the gene they are going to regulate). Concerning robustness in this context, there is an interpretational issue. If, at a particular point in parameter space, burst size (frequency) changes in response to changes in certain parameters e.g. chromatin looping, levels of TFs, but burst frequency (size) remains unchanged, then biology can tune either burst size or frequency by effectuating different kinds of parameter changes. We can interpret this scenerio as if biology is making different choices to suit its needs. However, if there is a point (or region) in parameter space where, say, burst frequency does not change in response to any parameter change, then that would constitute a form of robustness, which is yet to be explored.

Measuring transcriptional bursts

Among different models describing the dynamics of transcription, the simplest one involves mRNA production at a constant rate and degradation at a rate proportional to the number of mRNA molecules produced. The corresponding steady-state distribution of mRNA copy numbers follows a Poisson distribution. On the other hand, for a two-state model, where the promoter switches between an "on" and "off" states, the model can be described as a telegraph model (Nicolas et al., 2017). The "on" and "off" state represent transcriptionally active and inactive state, respectively. Depending upon the resident time in the "on" and "off" state, the telegraph model can produce a variety of shapes of mRNA copy number distribution (Chubb et al., 2006; Munsky et al., 2012; Peccoud and Ycart, 1995; Shahrezaei and Swain, 2008). When 'on' states are very short, the shape of the distribution becomes super-Poissonian, with long tail and high variance (Nicolas et al., 2017, Fig. 1(C)). Alternatively, when cells spend a long time in the 'on' or 'off' state, the distribution becomes bimodal (Munsky et al., 2012). The findings above conclude that the shape of mRNA distribution provides information about the promoter dynamics, and the two-state model is a useful tool to study transcriptional bursts. Assuming the telegraph model for transcription, one can obtain the bursting parameters by estimating them from the data obtained by smFISH experiments (Raj et al., 2006) using the maximum likelihood approach (Dey et al., 2015). Alternatively, comparing the moments of the mRNA distribution with the model prediction, parameters can be estimated (Peccoud and Ycart, 1995). Thus, the telegraph model provides a simple conceptual framework to estimate bursting parameters that has profound implications in understanding the effect of genetic and environmental perturbations on transcriptional output. Examples include investigating the effect of histone modification on transcriptional bursting in embryonic stem cells (Kim and Marioni, 2013). Additionally, under the assumptions that the on-state duration is considerably shorter than that of the off-state and mRNA lifetime, and the burst size is large, the normalised

burst frequency and the size can be obtained directly from the mean and variance of the mRNA copy numbers alone (Dey et al., 2015; Kim and Marioni, 2013; Nicolas et al., 2017). The burst frequency is inversely proportional to the square of CV, and the burst size is proportional to the ratio between the mean number of transcript and burst frequency. The two-state model is widely accepted for its simplicity and ease of obtaining analytical expressions for bursting parameters and mRNA copy number distributions that fit well with the smFISH data. In reality, transcription involves multilevel complex regulatory mechanisms such as DNA looping or TFs binding. Capturing the whole process in its entirety is a challenging task, not only because there is a lack of modelling approaches to handle the underlying complexity, but because there is also a lack of single-cell experimental procedures that can capture information about all such complex regulatory mechanisms at the same time. In this regard, Schwabe et al. (2012) compared a few alternative gene regulation models taking into account some of these complex regulatory mechanisms. The authors observed that genes with complex multiprotein regulation could have peaked burst-size distributions unlike the geometric (long-tailed) one in a two-state model. Unlike the two-state model, a complex mechanism of gene expression results in a non-exponential waiting time between gene switching and transcription initiation, which further decreases the noise in mRNA copy numbers and burst size. Besides, the authors remarked that the same experimental data could be well-fitted by qualitatively different regulatory models, though the bursting statistics between the models are entirely different.

Overall, the discussion above leads to the fact that obtaining mRNA distributions is essential to understand the mechanisms of transcriptional regulation and associated noise in terms of bursting parameters. However, at the same time, we require a model that offers a compact representation of the transcription process along with different regulatory mechanisms described above. Therefore, in this chapter, we present a rule-based approach that not only offers a graph-based abstraction of transcription considering different regulatory mechanisms but also provides a direct simulation platform for that graph-based abstraction based on the language κ to obtain mRNA copy numbers corresponding to different gene regulation models. Before we provide a detailed description of the rule-based model, it is necessary to have an overview of the process of transcription, which we introduce in the next section.

4.2 Mechanisms of gene transcription

Gene transcription is the first step towards gene expression. The process involves unwinding of double-stranded DNA followed by rewinding and synthesis of mRNA molecules. In this section, we provide a short overview of three major stages of transcription: initiation,

elongation, and termination.

Initiation

Initiation phase is considered as the "most heavily regulated phase" (Friedman and Gelles, 2012) in transcription. In viruses, transcription initiation can take place via viral polymerase[4] without the aid of cofactors. In bacteria and eukaryotes, the polymerase is aided by TFs to recognise and bind to the promoter region[5]. In eukaryotes, recruitment of polymerase is facilitated via the *mediator* complex (a multiprotein complex having 26 subunits in mammals and 21 subunits in yeast (Allen and Taatjes, 2015)). A mediator complex acts as a communicator between TFs and polymerase (Allen and Taatjes, 2015; Myers and Kornberg, 2000). TFs together with the polymerase form PIC. The PIC then unwinds the DNA at the promoter region and produces nascent RNA transcript that stabilises the complex (Dangkulwanich et al., 2014). In a productive pathway, when the length of nascent RNA is 9-11 nt long, the polymerase enters elongation phase. In an abortive pathway, RNAP produces short transcripts while remaining in the promoter region resulting in a *paused* state (Duchi et al., 2016). There exist multiple theories regarding the mechanism of pausing during transcription initiation. Interested readers are referred to Lerner et al. (2017); Roberts (2014) for more details. As a rate-limiting step, pausing in transcription controls the rate, timing, and magnitude of the response of the transcription (Liu et al., 2015). Since in previous chapters we focused on signalling pathways, it is worth to mention that signalling pathways have been found to influence the pausing of polymerase (Liu et al., 2015), and thus control the output of transcription. In case of RNAPII mediated transcription, the pause state occurs after production of $\sim$ 20-60 nt long nascent RNA transcript, and with the help of cofactors like the DRB sensitivity inducing factor (DSIF) and the negative elongation factor (NELF). A general observation is that the pause in RNAPII is released via recruitment of the positive transcription elongation factor (p-TEFb), though the mechanism of recruitment of p-TEFb is unclear (Conaway and Conaway, 2013). Experimental evidence points towards the involvement of mediator complex in conjunction with p-TEFb in RNAPII pause release. A good review of mediator dependent pause release is provided in Conaway and Conaway (2013). Another theory associated with the pause release is via enhancer mediated chromatin loop formation (Meng and Bartholomew, 2018). p-TEFb is introduced to the promoter region via the formation of the loop to release RNAPII from the paused state. In this way, chromatin looping modulates the output of transcription. Experimental evidences can be found in Bartman et al. (2016) for the

[4] A polymerase is an enzyme that facilitates the synthesis of long-chain polymers and nucleic acids such as DNA and RNA.

[5] A promoter is a region in the upstream (towards 3' region of the anti-sense strand) of DNA that initiates the transcription. The promoter region is usually 100-1000 base pairs long

β-globin gene.

Elongation

During the elongation phase, mRNA transcripts are synthesised in the direction from 5' to 3' according to the coding strand of DNA as RNAP traversed the template strand of DNA from 3' to 5'. In eukaryotes, an elongating polymerase may face obstacles due to nucleosomes, which can be taken care of by transcription elongating factors such as TFIIS (Fitz et al., 2016). At the end of elongation and after the release of DNA template by a single RNAP, re-initiation of transcription can happen on the same template by the same RNAP (Dieci et al., 2013; Hahn, 2004).

Termination

Transcription termination takes place when the polymerase receives the signal for termination. The terminator sequence provides the signal for termination at the end of the gene which is being transcribed. For bacteria, after termination, the RNA transcript acts as a mature mRNA molecule without any further processing. For eukaryotes, the generated RNA transcript undergoes two-step modifications. At first, a 5' cap at the beginning and a 3' poly-A tail at the end of the RNA transcript are added. Next, non-coding intron parts are removed, and coding exon parts are joined together to form the mature mRNA molecule.

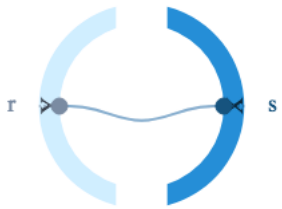
4.3 κ - a platform for rule-based modelling in molecular biology

It is evident from the aforementioned description that transcription involves multiple protein-protein, and protein-DNA interactions (TFs, TFs and promoter region), enzyme activities (RNAP), and phosphorylation reactions (p-TEFb for pause release). Usually these proteins have multiple sites that may undergo various modifications such as phosphorylations, methylation, and many more. This may lead to combinatorial explosion due to various modified forms that are difficult to write down following an ODE-based formalism for further analysis. In such scenarios, a rule-based formulation has been found to be efficient (Danos et al., 2007). The language κ was originally proposed to model protein-protein interactions and formally realized as a *sited graph* (Danos and Laneve, 2004). In such a sited graph, a protein is treated as a node or as an *agent* having multiple *sites* representing interfaces at which interactions/modifications take place (Boutillier et al., 2018), and a complex as a connected graph of such nodes or agents when proteins interact to each other

through their sites. The main idea of a rule-based language such as κ is to apply a specific transformation or a rule to an instance or a site graph having a *pattern* that matches the rule. For a reaction mixture comprised of several disconnected site graphs, *activity* of a rule (equivalent to propensity of a reaction to fire) is defined as the constant rate at which the rule triggers times the total number of distinguishable physical configurations generated upon application of that rule to the mixture (Boutillier et al., 2018). In the next three subsections, we introduce briefly the syntax of κ, the concept of concurrency in κ formalism, and the way κ rules are simulated using the Gillespie's SSA.

Syntax

Consider an agent X with a binding site as s. Now, $s[.]$ denotes that the site is free or unbound, and $s[n]$, where $n \in \mathbb{Z}_{>0}$, denotes that the site is bound. An expression $\mathrm{X}(s[1]), \mathrm{Y}(r[1])$ denotes that the agent X is bound at site s to the agent Y at site r. Visually the configuration can be realized as a contact map[6] :



From chemical reaction network to κ rules

In κ, rules are analogous to reactions in organic chemistry except the fact that rules codify observations irrespective of their biochemical relevance (Feret et al., 2009). Contrary to the ODE-based modelling, a rule-based approach offers a compact and transparent way to handle the combinatorial complexity of a chemical reaction network involving multiple proteins with multiple binding sites.

Example. *Consider the following set of chemical reactions (omitting rate constants for clarity):*

$$\begin{aligned} A &\leftrightarrow A^* \\ A + B &\leftrightarrow AB \\ AB &\leftrightarrow A^*B \\ A^* + B &\leftrightarrow A^*B, \end{aligned}$$

[6]Source: `https://tools.kappalanguage.org`

where A^* denotes the phosphorylated form of A. The above representation can be put into sentences in the following way:

C1. A can be phosphorylated and becomes A, even bound with B;*

C2. A can bind to B no matter its phosphorylation status, reversibly;

C3. Dephosphorylation of A, even bound with B.*

For the above set of chemical reactions, the ODE-based approach requires five species, A, A^*, B, AB, and A^*B to be listed. Whereas, due to the 'don't care, don't write' property of κ, the rule-based approach requires only two agents $A(x, z\{u, p\})$, $B(y)$, where x is the binding site for B, and the following two rules:

$$r_1 : A(x[.]), B(y[.]) \leftrightarrow A(x[1]), B(y[1]);$$
$$r_2 : A(z\{u\}[.]) \leftrightarrow A(z\{p\}[.]).$$

The rule r_1 fulfills the condition $C2$, and the rule r_2 fulfills the conditions $C1$ and $C3$.

Another important and perhaps the most significant property of κ or any rule-based language is the *concurrency*. An ODE-based modelling approach follows a global clock assuming that reaction events occur in a synchronized fashion. But the cell does not possess such clock meaning that reactions can happen in any sequence maintaining a partial order (some events can happen before, others can happen in any sequence) rather than a total order (Danos et al., 2007). Such causality of events is preserved in rule-based modelling approaches such as κ.

Example. *Here, we consider a series of phosphorylation-dephosphorylation events as described in (Danos et al., 2007) with notational modifications according to version 4.0 of κ.*

$$r_1 : Kin(a[.]), Tar(x[.]) \leftrightarrow Kin(a[1]), Tar(x[1]);$$
$$r_2 : Kin(a[1]), Tar(x\{u\}[1]) \rightarrow Kin(a[1]), Tar(x\{p\}[1]);$$
$$r_3 : Kin(a[.]), Tar(y[.]) \leftrightarrow Kin(a[1]), Tar(y[1]);$$
$$r_4 : Kin(a[1]), Tar(y\{u\}[1]) \rightarrow Kin(a[1]), Tar(y\{p\}[1]);$$
$$r_5 : Pho(a[.]), T(x[.]) \leftrightarrow Pho(a[1]), Tar(x[1]);$$
$$r_6 : Pho(a[1]), T(x\{p\}[1]) \rightarrow Pho(a[1]), Tar(x\{u\}[1]);$$
$$r_7 : Pho(a[.]), T(y[.]) \leftrightarrow Pho(a[1]), Tar(y[1]);$$
$$r_8 : Pho(a[1]), T(y\{p\}[1]) \rightarrow Pho(a[1]), Tar(y\{u\}[1]),$$

where agents Kin(a[.]), Pho(a[.]), and Tar(x{u, p}[.], y{u, p}[.]) represent the kinase, phosphatase, and target protein, respectively. The target protein Tar has two sites x and y which can be phosphorylated or dephosphorylated. For reversible rules, let r_i and r_i^{rev} denote the forward and the backward rule, respectively. Now, an event due to the rule r_6 can happen only after the event due to the rule r_5, and before the event due to the rule r_5^{rev}. This notion of logical precedence or causation defines a partial order on any sequence of events. Whereas, events due to r_1 and r_3 can be concurrent meaning that they can take place at the same time.

Gillespie's simulation for κ rules

The concept of Gillespie's SSA is based on constructing a continuous-time Markov chain (CTMC) structure of the process in concern. For a κ rule r in a particular state x of the reaction mixture, the activity $a(x, r)$ is defined as (Krivine et al., 2009):

$$a(x, r) = k_r[s_r, x] \text{ for } k_r \in \mathbb{R}_{>0},$$

where k_r is the intrinsic rate of the rule. s_r is the rule left hand side, and $[s_r, x]$ is the number of matches s_r has in x. The probability that r will be the next rule to be applied is given by $\frac{a(x,r)}{\sum_r a(x,r)}$. The time δt elapsed until the rule r is applied is given by $p(\delta t > T) = \exp(-\sum_r a(x, r) * T)$. The aforementioned dynamics is precisely the Gillespie's SSA as explained in Appendix 6.6.

4.4 κ-based approach for a gene transcription model

In Section 4.2, we described the mechanism of gene transcription briefly by introducing three major stages and few intermediate stages. Here, we represent RNAPII mediated gene transcription as a graph where a node is referred to as a *microstate*, and an edge is a transition between a pair of microstates. The concept of microstate is adapted from Ahsendorf et al. (2014) and defined as the snapshot of the transcription machinery at a particular instance of time during transcription. By transcription machinery, we indicate the set comprised of the gene, TFs, RNAPII, and mature mRNA molecules. An individual member of the transcription machinery is referred to as *agent* in this model. The existence of a particular microstate is constrained by a set of assumptions for the model. Moreover, an edge in that graph or a transition between two *valid* microstates (microstates according to the assumptions for the model) exists upon satisfying those assumptions. Assumptions are represented using logical expressions. Such representation facilitates the automatic generation of the κ rules corresponding to valid transitions. Later on, the rules are simulated using SSA to obtain the mRNA distribution and related statistics. Note that assumptions

can be relaxed or augmented according to the desired level of complexity in the model. The correctness of the rules is verified numerically for three simple variants of the model, against an alternative modelling approach based on the solution of the CME obtained using generating functions (Nam, 2018).

Model

The model for RNAPII mediated gene transcription, denoted by $\mathrm{M_{RNAPII}}$, is a 3-tuple,

$$\mathrm{M_{RNAPII}} = \{\Sigma, \Gamma, F : \Gamma \times \Gamma \to \mathbb{R}_{\geq 0}\},$$

where Σ is the set of agents, Γ is the set of valid microstates, and F is the transition function which takes a pair of microstates as argument and maps it to the set $\mathbb{R}_{\geq 0}$. A zero indicates an invalid transition.

Agents (Σ)

For $\mathrm{M_{RNAPII}}$, assuming one type of TF, we define the set Σ as:

$$\Sigma = \{\mathrm{G}, \mathrm{tf}, \mathrm{Pol}, \mathrm{En}\,, \psi\},$$

where G, tf, Pol, and En stand for gene, TF, RNAPII, and enhancer molecule respectively. The agent ψ records the number of mature mRNA molecules. Together with the respective sites, the agent set Σ can be rewritten in the following form

$$\Sigma = \{\mathrm{G}(f_1, f_2, \ldots, f_m, p_b, p_s, l_p, e_1, e_2, \ldots, e_N), \mathrm{tf}(\bar{f}), \mathrm{Pol}(\bar{p_b}, \bar{p_s}, \bar{e}_1, \bar{e}_2, \ldots, \bar{e}_N), \mathrm{En}(\bar{l_p}), \psi()\},$$

where $f_1, f_2, \ldots, f_m$ are designated to bind m copies of agent tf at site $\bar{f}$, RNAPII binding site p_b interacts with the corresponding site $\bar{p_b}$ in Pol, site p_s, when occupied, indicates a paused state, site l_p indicates the formation of the loop, and sites $e_1, e_2, \ldots, e_N$ are the indicators for N RNAPII molecules in the elongation phase and interact with the corresponding site $\bar{e}_1, \bar{e}_2, \ldots, \bar{e}_N$ in Pol respectively. As the agent ψ does not interact with any other agents there is no explicit site for it.

According to the notation of κ, identifier variable a in $A(x!a)$, takes any positive integer value. To construct logical expressions, we assume the identifier variable takes either a value of 0 or 1, to indicate whether the site is bound or not.

Set of valid microstates (Γ)

The generic form of a microstate μ can be represented as follows:

$$\begin{aligned}\mu =&\{\mathrm{G}\big(f_1!\, i_1, f_2!\, i_2, \ldots, f_m!\, i_m, p_b!\, a, p_s!\, b, l_p!\, c, e_1!\, \beta_1, e_2!\, \beta_2, \ldots, e_N!\, \beta_N\big),\\ &\mathrm{tf}(\bar{f}!\, i_1), \mathrm{tf}(\bar{f}!\, i_2), \ldots, \mathrm{tf}(\bar{f}!\, i_m)\,, \mathrm{Pol}\big(\bar{p_b}!\, a, \bar{p_s}!\, b, \bar{e}_1!\, \beta_1, \bar{e_2}!\, \beta_2, \ldots, \bar{e_N}!\ \beta_N),\ En(\bar{l_p}!\, c), \psi\}.\end{aligned} \tag{4.1}$$

Or, in a parametric form,

$$\mu = (i_1, i_2, \ldots, i_m, a, b, c, \beta_1, \beta_2, \ldots, \beta_N, \psi) \tag{4.2}$$

where $\{i_j\}_{j=1}^{m}, a, b, c$ and $\{\beta_k\}_{k=1}^{N}$ is either 0 or 1 indicating unbound or bound status of the corresponding sites, respectively. The agent ψ acts as counter variable and takes the value of any non-negative integer. From now onwards, we use the parametric form of μ as indicated by Equation (4.2). As the indicator variables are Boolean, we can construct logical expressions to describe microstates. For example, consider a microstate μ_h in which all TFs are bound to the promoter region and two molecules of RNAPII (indexed by 1 and 2) are elongating. In this setting, μ can be expressed through the following logical construct:

$$\left(\bigwedge_{k=1}^{m} i_k\right) \wedge \left(\sim a \wedge \sim b \wedge \sim c\right) \wedge \left(\sim \bigvee_{r=3}^{N} \beta_r\right) \wedge \beta_1 \wedge \beta_2, \tag{4.3}$$

where $\sim$ denotes the logical NOT. Therefore, validity of a microstate μ can be verified by evaluating the logical condition formed by the parameters or indicator variables under the following assumptions for our model $\mathrm{M}_{\mathrm{RNAPII}}$:

(A1) All TFs must remain bound to the promoter region during the stages of recruitment and pause of RNAPII and formation of the enhancer based chromatin loop. This assumption is according to the terminal recruitment strategy or all-or-none strategy of TF bindings as described in Estrada et al. (2016, Figure 3(A)).

(A2) RNAPII will either be on the initiation site or in the paused state.

(A3) RNAPII escapes from the paused state through enhancer mediated chromatin loop formation.

(A4) After escaping from the paused state, RNAPII enters the elongation phase. For a transcriptional process involving multiple copies of RNAPII, one copy of RNAPII can be in the recruitment stage while another one is elongating. Elongating copies of RNAPII maintain a sequential order while elongating. During the process of elongation, unless a new RNAPII is in the recruitment stage, TFs can unbind from the promoter region.

Assumptions (A1), (A2), and (A3) can be put together in the following logical form:

$$(\sim a\wedge \sim b\wedge \sim c)\vee \underbrace{\Bigg[\overbrace{\Bigg(\bigwedge_{j=1}^{m} i_j\Bigg)}^{\text{all-or-none strategy}}\wedge[(a\wedge \sim b\wedge \sim c)}_{\text{terminal recruitment}}\vee \overbrace{(\sim a\wedge b\wedge \sim c)}^{\text{pause}}\vee \underbrace{(\sim a\wedge b\wedge c)}_{\text{loop formation}}\Bigg], \quad (4.4)$$

and assumption (A4) takes the following logical form:

$$\underbrace{\Bigg[(\sim \beta_1\wedge \sim \beta_2\wedge,\ldots,\sim \beta_N)\vee(\beta_1\wedge \sim \beta_2\wedge,\ldots,\sim \beta_N)\vee,\ldots,\vee\Bigg(\bigwedge_{k=1}^{N}\beta_k\Bigg)\Bigg]}_{\text{elongation in a sequential order}}. \quad (4.5)$$

Conditions (4.4) and (4.5) together provide the final logical condition which must be satisfied by a microstate μ in order to qualify as a valid microstate:

$$\Bigg[(\sim a\wedge \sim b\wedge \sim c)\vee\Bigg[\Bigg(\bigwedge_{j=1}^{m} i_j\Bigg)\wedge[(a\wedge \sim b\wedge \sim c)\vee(\sim a\wedge b\wedge \sim c)\vee(\sim a\wedge b\wedge c)\Bigg]\Bigg]$$
$$\wedge\Bigg[(\sim \beta_1\wedge \sim \beta_2\wedge,\ldots,\sim \beta_N)\vee(\beta_1\wedge \sim \beta_2\wedge,\ldots,\sim \beta_N)\vee\ldots,\vee\Bigg(\bigwedge_{k=1}^{N}\beta_k\Bigg)\Bigg] \quad (4.6)$$

and for the minimal recruitment strategy (Estrada et al., 2016, Figure 3(B)), where at least one TF must be bound for RNAPII recruitment, the condition becomes:

$$\Bigg[(\sim a\wedge \sim b\wedge \sim c)\vee\Bigg[\overbrace{\Bigg(\bigvee_{j=1}^{m} i_j\Bigg)}^{\text{one-or-more strategy}}\wedge[(a\wedge \sim b\wedge \sim c)\vee(\sim a\wedge b\wedge \sim c)\vee(\sim a\wedge b\wedge c)\Bigg]\Bigg]$$
$$\wedge\Bigg[(\sim \beta_1\wedge \sim \beta_2\wedge,\ldots,\sim \beta_N)\vee(\beta_1\wedge \sim \beta_2\wedge,\ldots,\sim \beta_N)\vee\ldots,\vee\Bigg(\bigwedge_{k=1}^{N}\beta_k\Bigg)\Bigg] \quad (4.7)$$

Therefore, the set Γ contains all such microstates which satisfy either condition (4.6) or (4.7) depending on the recruitment strategy.

Example 1. *Consider a transcription process involving two copies of elongating RNAPII and two TF binding sites following a terminal recruitment strategy. Now, consider a microstate μ expressed in the following parametric form:*

$$\mu = (i_1, i_2, a, b, c, \beta_1, \beta_2) = (1, 0, 1, 0, 0, 1, 0).$$

The microstate describes a scenerio when one TF molecule is bound in the promoter region, an RNAPII molecule is recruited. No RNAPII is in the paused state and the chromatin

loop has not been formed yet. One RNAPII molecule is in the elongation phase. From the description it is evident that this is a clear violation of the terminal recruitment strategy as under this strategy no RNPAII can be recruited unless all the TFs are bound to the promoter region. As a consequence, the condition (4.6) *ouputs zero when evaluated using* μ*. Therefore, under the aforementioned settings,* μ *is not a valid microstate.*

Transition function

The transition function F between a pair of valid microstates (μ, μ^*) can be written in the following form:

$$F(\mu, \mu^*) \to \mathbb{R}_{\geq 0} \tag{4.8}$$

where

$$\begin{aligned}\mu =&(i_1, i_2, \ldots, i_m, a, b, c, \beta_1, \beta_2, \ldots, \beta_N, \psi)\\ \mu^* =&(i_1^*, i_2^*, \ldots, i_m^*, a^*, b^*, c^*, \beta_1^*, \beta_2^*, \ldots, \beta_N^*, \psi^*),\end{aligned}$$

and the transition takes place from μ to μ^*. We call a transition **valid** iff $F(\mu, \mu^*) > 0$ which, on the other hand, is possible only when one of the following conditions holds:

TFs binding and unbinding

During the process of binding/unbinding of TFs, the rest of the conditions such as RNAPII recruitment, pause, pause release, elongation, and mRNA count remain unchanged. Therefore a transition of this type will be a valid one if the following condition holds:

$$H(\mathbf{i}, \mathbf{i}^*) \wedge (a = a*) \wedge (b = b*) \wedge (c = c*) \wedge \left(\bigwedge_{j=1}^{N} (\beta_j = \beta_j^*) \right) \wedge (\psi = \psi^*), \tag{4.9}$$

where H is the Hamming distance between two vectors $\mathbf{i} = (i_1, i_2, \ldots, i_m)$, and $\mathbf{i}^* = (i_1^*, i_2^*, \ldots, i_m^*)$.

RNAPII binding and unbinding

In our model, we assume a terminal recruitment strategy for RNAPII binding. Therefore, all the TFs should remain bound at the promoter region to facilitate the binding and unbinding of RNAPII at the initiation site. Moreover, it is assumed that no other RNAPII is in the paused state. This assumption is in conjunction with the fact that a paused RNAPII can sterically hinder the recruitment of another RNAPII at the initiation site.

$$\left(\bigwedge_{j=1}^{m} i_j \right) \wedge \left(\bigwedge_{j=1}^{m} i_j^* \right) \wedge \overbrace{\sim a \wedge a^*}^{binding} \wedge \sim b \wedge \sim b^* \wedge \sim c \wedge \sim c^* \wedge \left(\bigwedge_{k=1}^{N} (\beta_k = \beta_k^*) \right) \wedge (\psi = \psi^*), \tag{4.10}$$

and for unbinding

$$\left(\bigwedge_{j=1}^{m} i_j\right) \wedge \left(\bigwedge_{j=1}^{m} i_j^*\right) \wedge \overbrace{a \wedge \sim a^*}^{\text{unbinding}} \wedge \sim b \wedge \sim b^* \wedge \sim c \wedge \sim c^* \wedge \left(\bigwedge_{k=1}^{N} (\beta_k = \beta_k^*)\right) \wedge (\psi = \psi^*). \tag{4.11}$$

Pause

During pause, the GTFs are assumed to remain bound in the promoter region.

$$\left(\bigwedge_{j=1}^{m} i_j\right) \wedge \left(\bigwedge_{j=1}^{m} i_j^*\right) \wedge a \wedge \sim a^* \wedge \sim b \wedge b^* \wedge \sim c \wedge \sim c^* \wedge \left(\bigwedge_{k=1}^{N} (\beta_k = \beta_k^*)\right) \wedge (\psi = \psi^*) \tag{4.12}$$

Enhancer mediated loop formation

Enhancer based loop formation is found to be associated with the paused polymerase (Ghavi-Helm et al., 2014). Experimental studies also suggest that the enhancer facilitates the pause release through loop formation (Liu et al., 2015). Therefore, it is reasonable to assume that during loop formation, RNAPII is still in the paused state, which gives the following condition:

$$\left(\bigwedge_{j=1}^{m} i_j\right) \wedge \left(\bigwedge_{j=1}^{m} i_j^*\right) \wedge \sim a \wedge \sim a^* \wedge b \wedge b^* \wedge \overbrace{\sim c \wedge c^*}^{\text{looping}} \wedge \left(\bigwedge_{k=1}^{N} (\beta_k = \beta_k^*)\right) \wedge (\psi = \psi^*), \tag{4.13}$$

and for unlooping

$$\left(\bigwedge_{j=1}^{m} i_j\right) \wedge \left(\bigwedge_{j=1}^{m} i_j^*\right) \wedge \sim a \wedge \sim a^* \wedge b \wedge b^* \wedge \overbrace{c \wedge \sim c^*}^{\text{unlooping}} \wedge \left(\bigwedge_{k=1}^{N} (\beta_k = \beta_k^*)\right) \wedge (\psi = \psi^*). \tag{4.14}$$

Elongation

Following an escape from the paused state, RNAPII enters into the phase of elongation. The elongation rate can be sequence dependent, and hence can be non-uniform. For simplicity, here we do not consider any velocity for elongation, though we briefly introduce the concept in Appendix 6.8. Depending upon the number of elongating RNAPII, there can be two different scenarios:

1. Only one RNAPII is elongating
2. More than one RNAPIIs are elongating simultaneously. Experimental evidences can be found in Klumpp and Hwa (2008); Padovan-Merhar et al. (2015); Senecal et al. (2014); Xu et al. (2015).

For multiple RNAPIIs, we have the following condition for the elongation:

$$\left(\bigwedge_{r=1}^{m} i_r\right) \wedge \left(\bigwedge_{r=1}^{m} i_r^*\right) \wedge \sim a \wedge \sim a^* \wedge b \wedge \sim b^* \wedge c \wedge \sim c^* \wedge \sim \beta_k \wedge \beta_k^* \wedge \left(\bigwedge_{j \neq k}^{N} \beta_j = \beta_j^*\right) \wedge (\psi = \psi^*), \tag{4.15}$$

where $k = \arg\min_j \{\beta_j = 0\}$. For $N = 1$, the condition (4.15) reduces to

$$\left(\bigwedge_{j=1}^{m} i_j\right) \wedge \left(\bigwedge_{j=1}^{m} i_j^*\right) \wedge \sim a \wedge \sim a^* \wedge b \wedge \sim b^* \wedge c \wedge \sim c^* \wedge \sim \beta_1 \wedge \beta_1^* \wedge (\psi = \psi^*). \tag{4.16}$$

When we have multiple copies of elongating RNAPII, say N, the transition from pause release to elongation must follow a sequential order. For example, a copy of RNAPII having an index $0 \leq k < N, k \in \mathbb{Z}$, is elongating. Therefore, the next copy of elongating RNAPII will be indexed by $k+1$ even there are $N - k + 1$ possibilities.

Termination and production of mature mRNA

During elongation, TFs can bind and unbind from the promoter region. A transition to termination is marked by an increase of the number of mRNA molecules by 1. Moreover, like the transition to the elongation, the order of termination is sequential. For example, if there are k copies of RNAPIIs currently elongating, the k^{th} copy of RNAPII will terminate first followed by the $(k-1)^{th}$ copy of RNAPII and so:

$$\left(\bigwedge_{r=1}^{m} (i_r = i_r^*)\right) \wedge (a = a^*) \wedge (b = b^*) \wedge (c = c^*) \wedge \beta_k \wedge \sim \beta_k^* \wedge \left(\bigwedge_{j \neq k}^{N} \beta_j = \beta_j^*\right) \wedge (\psi^* = \psi + 1), \tag{4.17}$$

where $k = \arg\max_j \{\beta_j = 1\}$.

mRNA degradation

mRNA degradation is marked by the decrease of the number of mature mRNA molecules by 1. Rest of the system will remain unchanged:

$$\left(\bigwedge_{j=1}^{m} (i_j = i_j^*)\right) \wedge (a = a^*) \wedge (b = b^*) \wedge (c = c^*) \wedge \left(\bigwedge_{k=1}^{N} (\beta_k = \beta_k^*)\right) \wedge (\psi^* = \psi - 1) \wedge (\psi >= 1) \tag{4.18}$$

Automatic generation of κ rules

Once we fix the model, the next task is to automate the generation of κ- rules with the language specific syntax that can be directly fed to the κ simulator to generate the samples

paths using SSA.

Example. *Let us consider a model where there are two binding sites for the same TF. There is no explicit binding site for RNAPII. For a particular binding configuration, say when all the binding sites are occupied, transcription can proceed and an mRNA molecule is produced. For this model, the parametric form of the microstate is*

$$\mu = (i_1, i_2, \psi).$$

Since there is no explicit binding site for RNAPII, the set of valid microstates will not be depending on the steps that depend on RNAPII. Therefore the system will transit in a stochastic way between four different states (not considering degradation as a state):

$$\mu_1 = (0, 0, \psi), \mu_2 = (0, 1, \psi), \mu_3 = (1, 0, \psi), \mu_4 = (1, 1, \psi + 1)$$

Following the κ syntax of version 4.0, we have the corresponding rules:

$r_1 : G(f_1[.], f_2[.]),\ tf(\bar{f}[.]) \leftrightarrow G(a[1], b[.]),\ tf(\bar{f}[1])\ @\ a_{00\to10},\ a_{10\to00}$
$r_2 : G(f_1[.], f_2[.]),\ tf(\bar{f}[.]) \leftrightarrow G(a[.], b[1]),\ tf(\bar{f}[1])\ @\ a_{00\to01},\ a_{01\to00}$
$r_3 : G(f_1[1], f_2[.]),\ tf(\bar{f}[1]),\ tf(\bar{f}[.]) \leftrightarrow G(f_1[1], f_2[2]),\ tf(\bar{f}[1]),\ tf(\bar{f}[2])\ @\ a_{10\to11},\ a_{11\to10}$
$r_4 : G(f_1[.], f_2[2]),\ tf(\bar{f}[.]),\ tf(\bar{f}[2]) \leftrightarrow G(f_[1], f_2[2]),\ tf(\bar{f}[1]),\ tf(\bar{f}[2])\ @\ a_{01\to11},\ a_{11\to01}$

$a_{x\to y}$ indicates the rate of transition from state x to state y encoded as binary strings. A '1' implies an occupied position. Now, adding mRNA production and degradation stages, we have additionally the following two rules:

$r_5 : G(f_1[1], f_2[2]),\ tf(\bar{f}[1]) \to G(f_1[1], f_2[2]),\ tf(\bar{f}[1]),\ tf(\bar{f}[2]), \psi()\ @\ k_{init}$
$r_6 : \psi() -\ @\ k_{degrade},$

where k_{init} and $k_{degrade}$ are transcription initiation rate and mRNA degradation rate, respectively.

These rules along with the declaration and initiation of agents and variables, can now be simulated using SSA to obtain the desired quantity declared as observable, say distribution of mRNA copy numbers at steady state (see Figure 4.1).

4.5 Numerical examples

We consider three alternative models of transcription, and for each model we numerically compare mean and variance of mRNA copy numbers at steady state obtained using the κ-based modelling approach to the one obtained using the CME based approach described

in Nam (2018). In this way, we cross validate the κ rules that are generated using an in-house software pipeline written in Python.

Model A

In this model variant, we assume two binding sites of the TF at the promoter region of the DNA molecule. In addition, we consider that the initiation of transcription process is based on the previously discussed all-or-none strategy i.e. transcription will be initiated only when all the TF binding sites are occupied. mRNA creation and degradation take place at a constant rate of 1. In this setting, the system makes transitions among four states (omitting the agent $\psi()$ that keeps track of mRNA copy numbers):

$$\mu_1 = (0,0), \mu_2 = (0,1), \mu_3 = (1,0), \mu_4 = (1,1),$$

where 1 and 0 denote occupancy and non-occupancy of the binding site, respectively. When the system is in state μ_4, it produces an mRNA molecule.

Model B

Model B is the same as model A except that the former has one more binding site. Hence, the number of states will be eight.

$$\mu_1 = (0,0,0), \mu_2 = (0,0,1), \mu_3 = (0,1,0), \mu_4 = (0,1,1),$$
$$\mu_5 = (1,0,0), \mu_6 = (1,0,1), \mu_7 = (1,1,0), \mu_8 = (1,1,1),$$

and the transcription is initiated when the system is at state $\mu_8 = (1,1,1)$.

Model C

In this model variant, we assume an explicit binding site for the RNAPII molecule denoted by Pol. In the parametric form of μ, the occupancy and non-occupancy of site Pol are denoted by 1 and 0, respectively. Furthermore, we also assume that the transcription starts once the RNAPII is bound, and either all the TF binding sites are occupied or only the first binding site is occupied. For this case also, the number of states will be eight, but the initiation states are $\{(1,1,\text{Pol}=1),(1,0,\text{Pol}=1)\}$.

Simulation results

For the κ-based approach, the expectation and variance at steady state are calculated from an ensemble of 1000 SSA realizations for 5, 10, 100, and 1000 TF molecules. For all the

model variants, we assume the value of $A_v V$ (Avogadro constant times the volume of the cell V) is *unity*. Therefore, the stochastic rate constant remains the same as its deterministic counterpart [7]. Figures 4.2(a)–4.2(c) summarise the results for numerical validation of the κ rules of the aforementioned gene regulatory models.

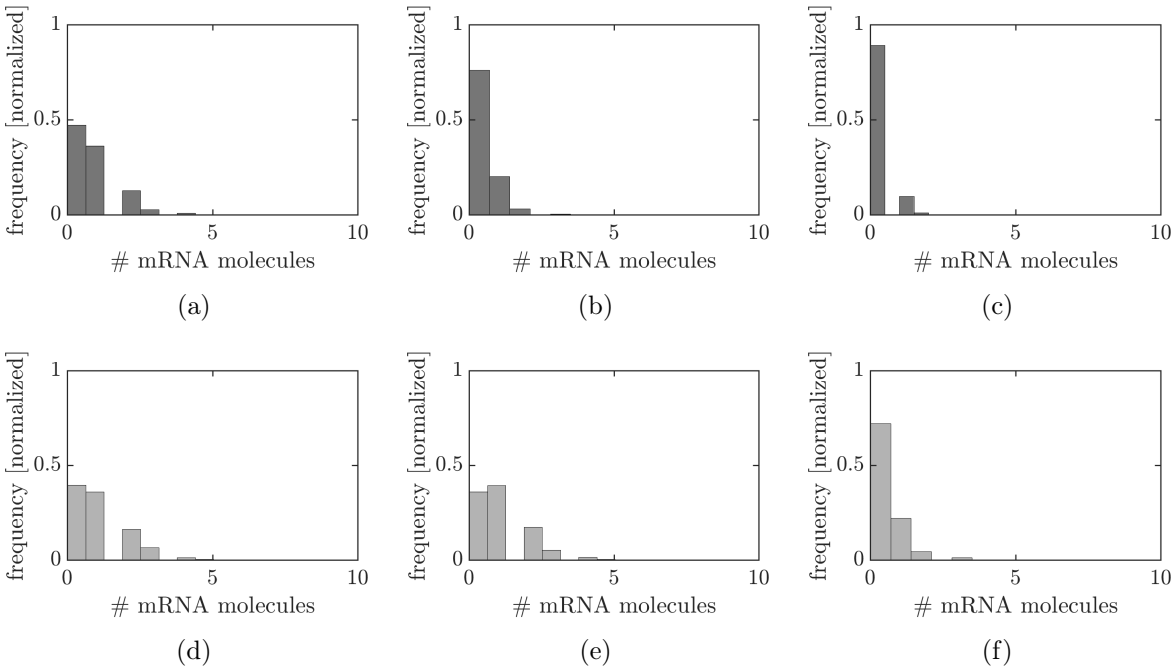


Fig. 4.1. Distribution of mRNA copy numbers at steady state. Distribution of mRNA copy numbers at steady state are plotted for model A,B, and C respectively (from left to right). The number of TF molecules are 5 (a)-(c), and 1000 (d)-(f), respectively. Each distribution corresponds to 1000 SSA realizations.

[7] Let α be the stochastic rate constant and η be the concentration-based or deterministic rate constant. Then $\eta = \alpha * (AV)^{a-1}$, where a is the arity of the reaction, for example, $a = 2$ for a bimolecular reaction.

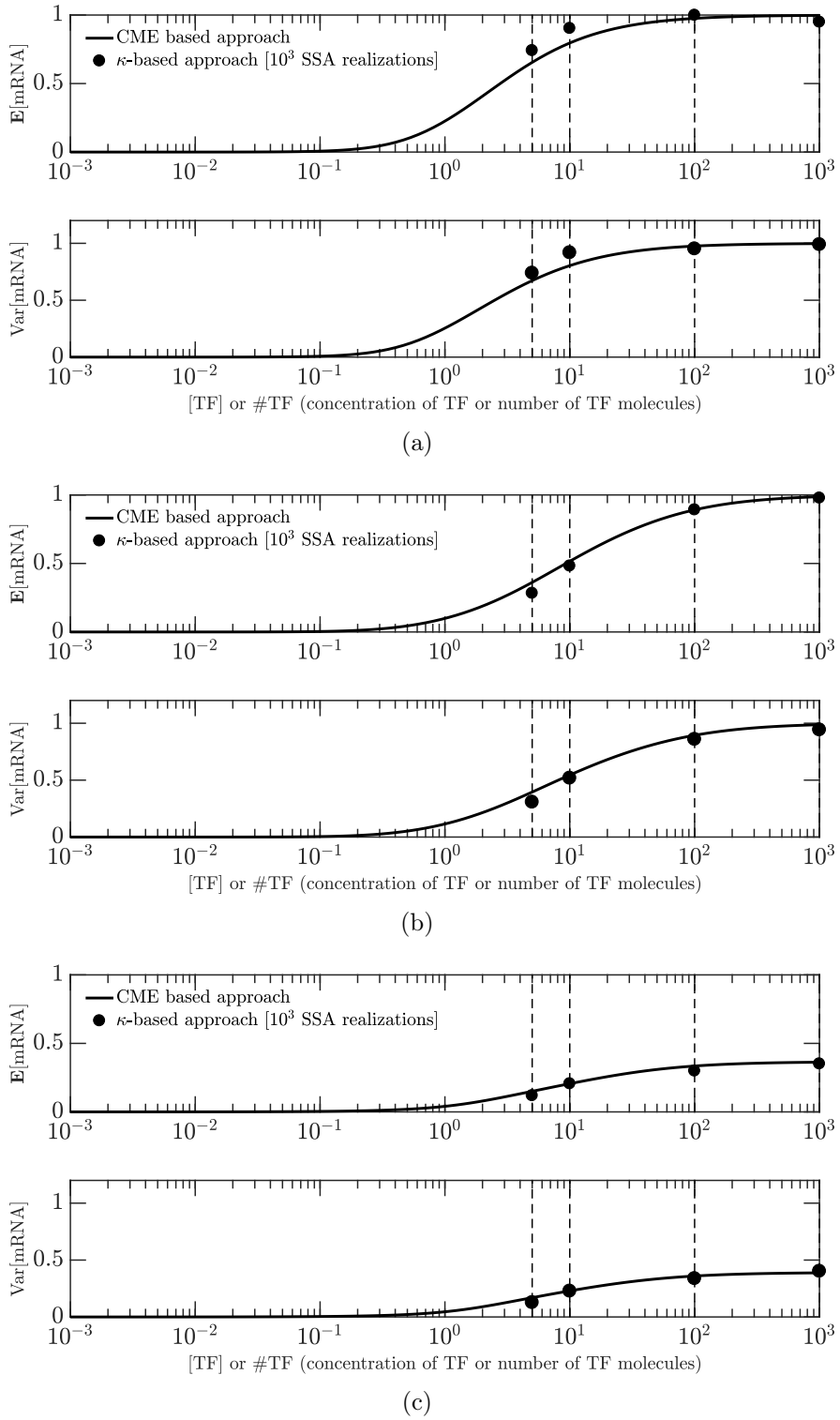


Fig. 4.2. Numerical validation of κ rules. For model variants (a) A, (b) B, and (c) C, steady state expectation (**E**[mRNA]) and variance (Var[mRNA]) of mRNA copy numbers across a range of concentrations of TFs are plotted using the CME based approach (Nam, 2018) as indicated by the continuous line. The output is compared against the one that is obtained using κ based approach as discussed previously. The comparison is made at TF number $5, 10, 100$, and 1000. The transcription initiation rate and the mRNA degradation rate is 1 for all three models. For other reaction rate constants, readers are referred to Nam (2018).

4.6 Summary and discussion

The goal of this chapter was to understand the mechanisms of robustness in gene expression, in particular, at the level of transcription. As the transcription occurs in bursts, which is the primary source of variability in gene expression, the goal then reduces to understanding how different regulatory mechanisms associated with transcription such as chromatin looping, availability of TFs, histone modifications, nucleosome occupancy, number of *cis*-regulatory elements, modulate the bursting parameters, thus affecting robustness. We understood that, for a simple description of transcription using a two-state model, estimating bursting parameters is straightforward, and can be obtained directly from the mRNA copy numbers. For complex regulatory models, this estimation is not straightforward. In those cases, we have to rely on mRNA copy numbers to understand the dynamics of transcription. In this chapter, we presented a graph-based formalism to the process of transcription that accommodates a subset of complex regulatory mechanisms mentioned above, though the formalism is flexible enough to include other regulatory mechanisms as well. Afterwards, we described how such graphical representation could be transformed into executable κ rules to obtain mRNA copy numbers under different regulatory assumptions (see Figure 4.1 for example) combining the mechanisms mentioned above. We validated the output of the rule-based model against the output of an alternative model (see Figure 4.2) based on the CME for three regulatory models of varying promoter complexity Nam (2018). Validation of the regulatory models opens up future possibilities to obtain and analyse mRNA copy number distributions under various regulatory assumptions and thereby facilitates a precise understanding of robustness, its mechanisms, and consequences in the context of gene expression.

5 Conclusion

This chapter concludes the thesis by summarising the main findings of all the preceding chapters, discussing general aspects, and presenting an outlook on a few potential future directions for this research.

5.1 Summary

Chapter 2 presented deterministic modelling approaches based on ODEs and simple mass-action kinetics to investigate whether a cascade of two double PD cycles (model C) exhibits robust input-output behaviour upon input perturbation compared to that of a single (model A) and a double PD cycles (model B). Through local sensitivity and variance-based analysis, we observed that at steady state, the output of model C is less sensitive to the input variation compared to that of model A and B (Figures 2.2–2.4). Furthermore, for a sinusoidal input, we observed that model C acts as an efficient low-pass filter compared to that of model A and B. Figure 2.5 and Bode plots (Figure 2.6) illustrate the fact.

In Chapter 3, we excluded the model of a single PD cycle and compared the input-output behaviour of the double PD cycle (renamed as model A) and the cascade of two double PD cycles (renamed as model B) for a stochastic environment with explicit addition of kinase and phosphatase molecules to the models. Moreover, we considered two-step enzyme kinetics to model all the chemical reactions. Additionally, we considered CV as the measurement for intrinsic noise. Through numerical approaches based on SSA, first, we observed that at steady state, dynamics of model B reduced the CV across a range of kinase molecules compared to that of model A (Figure 3.3). We noticed a further reduction in the CV upon increasing the sequestration of the terminal molecule ppX of model B (Figure 3.4). Based on these preliminary observations, we concluded that sequestration dynamics of ppX is responsible for such a reduction in the CV (Figure 3.2). However, the actual mechanism was still unknown at this point. Further analysis based on individual sample paths showed that the upstream module of model B dynamically adapts to the changes in the downstream module via a sequestration based retroactive effect (Figure 3.8). We referred to such mechanism as dynamic sequestration. Later on, we found the biological context of dynamic sequestration by adapting the parameters of the models to a biologically feasible range. On the contrary to non-biological context, observation in the biological

context revealed an ambivalent role of dynamic sequestration (Figure 3.10(a)). Upon dose-response analysis (Figure 3.11), we finally concluded that the ambivalent role of dynamic sequestration depends on the operating regimes of the output of the respective models.

Chapter 4 presented the state-of-the-art of an ongoing collaborative project which primarily aims at understanding the mechanisms of robustness in stochastic gene expression that is quantified as transcriptional bursts and mRNA copy numbers. To achieve the goal, as a first step, we introduced a rule-based modelling approach for a model of transcription based on the κ platform. The model of transcription has an underlying graphical structure, where a node signifies a microstate, and an edge signifies the transition between a pair of microstates. Afterwards, we introduced a set of logical conditions that define the edges in the aforementioned graph of microstates. The graphical abstraction was transformed into a set of executable κ-rules, which were then simulated using the SSA to obtain the mRNA copy numbers. Finally, using numerical simulations, we verified the correctness of the executable rules on three gene regulation models (Figure 4.2), against an alternative CME based approach described in (Nam, 2018).

5.2 Discussion

The design goal for technical systems is to create structures that remain operational under a wide range of adverse environmental conditions. Therefore, robustness is an explicit theme for technical designs. However, for biological systems the theme of robustness is implicit, meaning that biologists have long understood phenomena such as thermoregulation in homeothermic organisms across a wide range of ambient temperatures (Hammerstein et al., 2006) that resonates the idea of robustness. In spite of such implicitness, the study of robustness in biology is essential because it is the key to understand evolution (Kitano, 2007) as we envisioned in this thesis. After all,

Evolution is a light which illuminates all facts, a curve that all lines must follow.

– Pierre Teilhard de Chardin *p. 219 of The Phenomenon of Man*

5.3 Outlook

In this section, we discuss a few potential future directions of this thesis.

Non-linear frequency analysis

In Chapter 2, we linearized non-linear models around equilibrium points to obtain Bode magnitude plots for frequency analysis. At this point, the question arises whether the

notion of a transfer function exists for non-linear systems as well. The answer is yes. When the non-linear system is described in the time domain using Volterra functional series, one can realise the transfer function for that non-linear system (Billings and Zhang, 1994; Volterra, 1930). A rigorous mathematical formulation of such a transfer function for two non-linear systems in cascade has been presented in Barrett (1963); Kielkiewicz (1970). Analysis of non-linear transfer functions was unexplored for many years. Non-linear transfer functions are multivariate in nature even for single input/output systems. For this reason, the transfer function is difficult to analyse, and the interpretation of response characteristics for systems becomes complicated (Zhang and Billings, 1993). Moreover, as a consequence of a Volterra functional polynomial representation, the main transfer function for a non-linear system is made up of a sequence of transfer functions instead of one transfer function as in the linear case. Fortunately, for most of the non-linear systems, the dynamics is dominated by first, second, and third order transfer functions (Zhang and Billings, 1993). Nevertheless, analysing transfer functions for non-linear systems may open up some essential non-linear phenomena that are yet to be reported.

Enzyme processivity and sequestration effect

In Chapter 3, we assumed a distributive mechanism for a two-step enzymatic process. A distributive mechanism accounts for sigmoidal stimulus/response curve for the MAPK pathway in a *Xenopus oocyte* system (Huang and Ferrell, 1996). Ferrell and Bhatt (1997) provides mechanistic details on how the distributive mechanism is responsible for producing such a response. Recently, it has been found that the same MAPK pathway exhibits a graded stimulus/response curve in mammalian cells (Aoki et al., 2013, 2011). The authors found out that molecular crowding[1] is responsible for transforming a sigmoidal response to a graded response. Therefore, it will be interesting to observe the behaviour of dynamic sequestration when the enzyme kinetics is processive (see Appendix 6.7).

Integrating the effect of signalling pathways in regulation of gene expression

Signalling pathways and regulation of gene expression is often studied separately due to complexity and non-linearity in both the systems (den Breems et al., 2014). But, treating them in isolation will not be fruitful to understand fully the underlying mechanism of robustness in transcriptional regulation as signalling pathways such as the MAPK targets transcription factors, co-regulators, and chromatin proteins, to regulate DNA binding,

[1]The intracellular environment is highly crowded with different biomolecules. When biomolecules are in high concentration (occupying 20-30% of cellular volume), they tend to attract each other. Such physicochemical phenomenon is called molecular crowding (Cho and Kim, 2012).

protein stability, and cellular localization (den Breems et al., 2014; Whitmarsh, 2007; Yang et al., 2003), hence play a significant role in influencing the process of transcription. Therefore, integrating the effect of signalling cascades will be a potential future direction of this research.

6 Appendix

6.1 Sensitivity analysis

According to Saltelli et al. (2004), the definition of sensitivity analysis is the following: *The study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input.* Therefore, the simplest way to quantify sensitivity (s) is

$$s = \frac{\Delta \text{output}}{\Delta \text{input}},$$

where Δ represents the change in the corresponding quantities. In order to introduce the concept of sensitivity as adapted from Varma et al. (2005), let us consider the following dynamics of a system described by a single real variable λ as

$$\dot{\lambda} = \Phi(\lambda, \delta, t);\ \lambda(t = 0) = \lambda_0, \tag{6.1}$$

where Φ is a continuously differentiable function in all its arguments, and δ is a scaler parameter. The continuity and differentiability of Φ ensures uniqueness of the local solution which is continuous, differentiable in time t and in parameter δ, and takes the following form

$$\lambda = \lambda(\delta, t). \tag{6.2}$$

Furthermore, a perturbation in one of the parameters in δ say $\Delta\delta$ leads to the solution:

$$\lambda^* = \lambda(\delta + \Delta\delta, t), \tag{6.3}$$

A fundamental assumption here is that the change $\delta \to \delta + \Delta\delta$, does not occur very fast with respect to time. Furthrmore, if $|\Delta\delta| \ll 1$, the perturbed solution can now be expanded into a Taylor series (that converges) up to the first order term as follows

$$\lambda(\delta + \Delta\delta, t) \approx \lambda(\delta, t) + \nabla_\delta \lambda(\delta, t) \cdot \Delta\delta. \tag{6.4}$$

By rearranging the approximation in (6.4), we obtain the following linear approximation of $\Delta\lambda$

$$\Delta\lambda = \lambda(\delta + \Delta\delta, t) - \lambda(\delta, t) \approx \nabla_\delta \lambda(\delta, t) \cdot \Delta\delta. \tag{6.5}$$

Output sensitivity with respect to the input parameter δ can now be expressed as

$$S_\lambda(\delta) = \lim_{\Delta\delta\to 0} \frac{\Delta\lambda}{\Delta\delta} \approx \nabla_\delta \lambda(\delta, t). \tag{6.6}$$

A major drawback of Equation (6.6) is that, in practical scenarios, it may not be dimensionless and hence provides a wrong measure. For example, consider λ has a unit of concentration and δ_j has a unit of concentration·time^{-1}, then $S(\lambda, t)$ will have unit of time and provides different measures depending on the choice of timescale. This is certainly an indication that $S(\lambda, t)$ is not an appropriate measure in general. In addition, a dimensionless $S(\lambda, t)$ cannot be compared directly with another $S(\lambda, t)$ having a dimension. Finally, significance of $\Delta\lambda$ may vary depending upon the actual value of λ relative to $\Delta\lambda$. Therefore, to alleviate the problems mentioned previously, a dimensionless equivalent of $S(\lambda, t)$, called the *normalized local sensitivity*, has been introduced in the following way (ommiting the arguments of λ for clarity):

$$s_\lambda(\delta) = \frac{\delta}{\lambda}\nabla_\delta\lambda = \frac{\partial\lambda}{\lambda}\cdot\frac{\delta}{\partial\delta} = \frac{\partial \ln\lambda}{\partial \ln\delta} = \nabla_{\ln\delta} \ln\lambda. \tag{6.7}$$

6.2 Steady state expressions for models with proteins having more than two phosphorylation sites

Recall our general model class for multisite phosphorylation for a single protein X as depicted in Equation (2.1). Assuming input u as the only kinase, for the fully phosphorylated protein, we can obtain the steady state expression $\bar{x}_m$ in the following manner (omitting the superscript X on the rate constants in Equation (2.1)),

$$\frac{\bar{x}_m}{\bar{x}_{m-1}} = \frac{k_m u}{k_{-m}} \implies \bar{x}_m = \frac{k_m u}{k_{-m}}\bar{x}_{m-1}.$$

Thus by recursion, $\bar{x}_m$ can be expressed in terms of $\bar{x}_1$,

$$\bar{x}_m = \prod_{j=2}^{m} \frac{k_j u}{k_{-(j)}}\bar{x}_1 \tag{6.2.1}$$

Finally, the task is to obtain an explicit expression for $\bar{x}_1$ in terms of u and the rate constants. Using

$$\dot{x}_1 = k_1 u\left(1 - \sum_{j=1}^{m} x_j\right) + k_{-2}x_2 - k_{-1}x_1 - k_2 u x_1$$

for the steady state we get,

$$
\begin{aligned}
& k_1 u \left(1 - \sum_{j=1}^{m} \bar{x}_j\right) + k_{-2}\bar{x}_2 = k_{-1}\bar{x}_1 + k_2 u \bar{x}_1 \\
\implies & k_1 u \left(1 - \sum_{j=2}^{m} \bar{x}_j\right) + k_{-2}\bar{x}_2 = (k_{-1} + k_2 u + k_1 u)\bar{x}_1 \\
\implies & k_1 u \left(1 - \sum_{j=2}^{m} \bar{x}_j\right) + k_{-2}\left(\frac{k_2 u}{k_{-2}}\bar{x}_1\right) = (k_{-1} + k_2 u + k_1 u)\bar{x}_1 \\
\implies & k_1 u \left(1 - \sum_{j=2}^{m} \bar{x}_j\right) = (k_{-1} + k_1 u)\bar{x}_1 \\
\implies & k_1 u = (k_{-1} + k_1 u)\bar{x}_1 + k_1 u \sum_{j=2}^{m} \bar{x}_j \\
\implies & k_1 u = (k_{-1} + k_1 u)\bar{x}_1 + k_1 u \sum_{j=2}^{m} \prod_{n=2}^{j} \frac{k_n u}{k_{-n}} \bar{x}_1 \ \text{ [using Equation (6.2.1)]} \\
\implies & \bar{x}_1 = \frac{k_1 u}{\left(k_{-1} + k_1 u + k_1 u \sum_{j=2}^{m} u^j \prod_{n=2}^{j} \frac{k_n}{k_{-n}}\right)}
\end{aligned}
$$

For $m = 2$ we arrive at the following expression for $\bar{x}_1$,

$$\bar{x}_1 = \frac{k_1 u}{\left(k_{-1} + k_1 u + k_1 u \cdot \frac{k_2 u}{k_{-2}}\right)},$$

as described in Equation (2.8). In addition, for an input u, the expression for the normalized local output sensitivity coefficient for a single protein X with $m > 0$ phosphorylation sites takes the following form:

$$s_{y^*}(u) = \nabla_{\ln u} \ln \bar{x}_m(u) = \nabla_{\ln u} \ln \bar{x}_1(u) + m - 1. \tag{6.2.2}$$

Figure 6.1(a) demonstrates numerical verification of Equation (6.2.2) for triple phosphorylation.

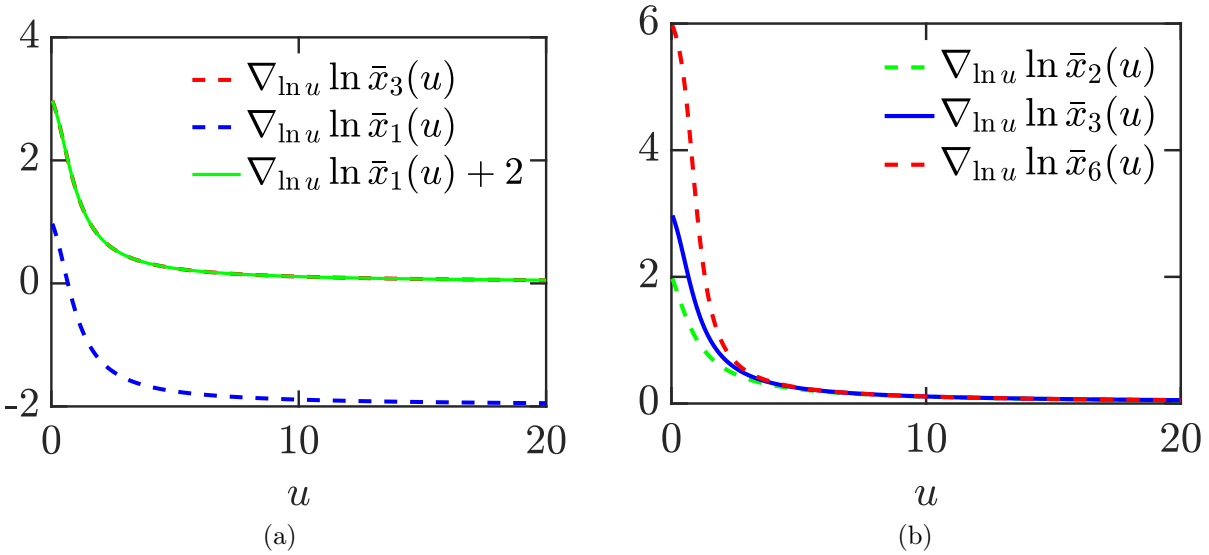


Fig. 6.2.1. Local sensitivities for multisite phosphorylation. (a) Numerical verification of Equation (6.2.2) for triple phosphorylation. Dashed blue and red lines represent numerically obtained values and the solid green line represents analytically calculated value using Equation (6.2.2) (b) Comparison of output sensitivities for $m = 2, 3$ and 6. All the rate parameters are set to 1.

6.3 Linear time invariant systems

In this section we introduce the concept of state-space representation for a linear time invariant (LTI) system and linearization of a non-linear dynamic system. Materials presented here are adapted from Hespanha (2018); Williams et al. (2007).

State-space representation for an LTI system

The state-equation for an LTI system takes the following general form

$$\begin{aligned} \dot{\mathbf{x}} &= A\mathbf{x} + B\mathbf{u}, \;\; x(t) \in \mathbb{R}^n, \;\; \mathbf{u} \in \mathbb{R}^m \\ \mathbf{y} &= C\mathbf{x} + D\mathbf{u}, \;\; \mathbf{y} \in \mathbb{R}^l, 1 \leq l \leq n \end{aligned} \tag{6.3.1}$$

where A is the state matrix of dimension $n \times n$, B is the *input-to-state* matrix of dimension $n \times m$, C is the *state-to-output* matrix of dimension $l \times n$ and D is the *feed-through* matrix of dimension $l \times m$, and all of them are constant matrices. In addition, $\mathbf{x}$ is the n-dimensional *state vector*, $\mathbf{u}$ and $\mathbf{y}$ are the m-dimensional *input vector* and l-dimensional *output vector* respectively. The initial state of the system is defined by the vector $\mathbf{x}(t_0) = \mathbf{x}_0 \in \mathbb{R}^n$.

Transfer function

For an LTI system described by Equation (6.3.1), the *transfer function* $H(s)$, where $s = \sigma + j\omega,\ \sigma, \omega \in \mathbb{R}$, is the ratio between the system's output and input in the Laplace domain when all the initial conditions are assumed to be zero.

Derivation

Applying Laplace transform on both the sides of the Equation (6.3.1) we get,

$$sX(s) = AX(s) + BU(s)$$
$$Y(s) = CX(s) + DU(s),$$

where s is the complex Laplace variable. Now,

$$(sI - A)X(s) = BU(s) \implies X(s) = (sI - A)^{-1}BU(s).$$

Furthermore, replacing $X(s)$ in $Y(s) = CX(s) + DU(s)$ with $(sI - A)^{-1}BU(s)$, we obtain

$$Y(s) = C(sI - A)^{-1}BU(s) + DU(s).$$

From this point, it is straightforward to have the expression for the transfer function:

$$H(s) = \frac{Y(s)}{U(s)} = C(sI - A)^{-1}B + D.$$

In case we have no feed-through matrix D, the above expression becomes

$$H(s) = C(sI - A)^{-1}B.$$

Remark. *A transfer function contains all the information regarding the order, type and frequency response of a control system. The frequency response can be analysed through Bode plots which is discussed in the main text. Although the terms frequency response and transfer function are closely related, the former is a representation of the input-output relationship in the Fourier domain* ($s = \sigma + j\omega,\ j = \sqrt{-1}$).

Linearization of the non-linear dynamic system

Most of the biological systems are non-linear in nature. Unlike a linear system, we cannot apply additivity and homogeneity transformations [1] for a non-linear dynamic system. But, we can linearize the non-linear system around a small neighborhood of the equilibrium points (the points where the differential equation vanishes), and when the equilibrium points are

[1] A function $f(x)$ is a *linear map* if it satisfies the following two properties:

- Additivity: $f(x + y) = f(x) + f(y)$
- Homogeneity: $f(\beta\, x) = \beta\, f(x)$ for β is a constant

hyperbolic (do not have any center manifolds), then the dynamics of the linearized system qualitatively approximate the dynamics of the non-linear system within that neighborhood. Suppose, a non-linear dynamic system is represented by the following state-space equation:

$$\begin{aligned} \dot{\mathbf{x}} &= f(\mathbf{x}, \mathbf{u}),\ \mathbf{x} \in \mathbb{R}^n, \mathbf{u} \in \mathbb{R}^m \\ \mathbf{y} &= g(\mathbf{x}),\ \mathbf{y} \in \mathbb{R}^l, 1 \leq l \leq m. \end{aligned} \tag{6.3.2}$$

When the functions f and g are at least continuous and differentiable within a small neighborhood of a point $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$, the functions can be well approximated by a properly defined linear system. Usually, the points $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ are considered to be *equilibrium points* at which f is zero. A special case is when $\bar{\mathbf{u}} = \mathbf{u}^*$ is a constant. In this scenerio $\bar{\mathbf{x}}$ serves as equlibrium point that satisfies $f(\bar{\mathbf{x}}, \mathbf{u}^*) = 0$. An expansion of f and g around $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ takes the following form:

$$\begin{aligned} f(\mathbf{x}, \mathbf{u}) &\approx f(\bar{\mathbf{x}}, \bar{\mathbf{u}}) + A(\mathbf{x} - \bar{\mathbf{x}}) + B(\mathbf{u} - \bar{\mathbf{u}}) \\ g(\mathbf{x}) &\approx g(\bar{\mathbf{x}}) + C(\mathbf{x} - \bar{\mathbf{x}}), \end{aligned} \tag{6.3.3}$$

where $A = \nabla_{\mathbf{x}} f$, $B = \nabla_{\mathbf{u}} f$, $C = \nabla_{\mathbf{x}} g$

Now, $f(\bar{\mathbf{x}}, \bar{\mathbf{u}}) = 0$ as $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ are equilibrium points. Furthermore, considering the deviations around $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$

$$\delta\mathbf{x} = \mathbf{x} - \bar{\mathbf{x}},\ \delta\mathbf{u} = \mathbf{u} - \bar{\mathbf{u}},\ \delta\mathbf{y} = \mathbf{y} - g(\bar{\mathbf{x}}),$$

we have the *linearized system* as

$$\begin{aligned} \delta\dot{\mathbf{x}} &= A\delta\mathbf{x} + B\delta\mathbf{u} \\ \delta\mathbf{y} &= C\delta\mathbf{x} \end{aligned} \tag{6.3.4}$$

6.4 Bode magnitude plot

A Bode magnitude plot for a LTI system is the plot of *magnitude* of the transfer function $H(s)$versus the frequency.

Poles and Zeros

The transfer function $H(s)$ is a rational function in the complex variable $s = \sigma + j\omega$ as follows:

$$H(s) = \frac{b_m s^m + b_{m-1} s^{m-1} +, \ldots, + b_1 s^1 + b_0}{a_n s^n + a_{n-1} s^{n-1} +, \ldots, + a_1 s^1 + a_0}.$$

It is convenient to factorize the polynomials and write down in the following format

$$H(s) = \frac{N(s)}{D(s)} = K\frac{(s-z_1)(s-z_2)\dots(s-z_{m-1})(s-z_m)}{(s-p_1)(s-p_2\dots(s-p_{n-1})(s-p_n)},$$

where $N(s)$ and $D(s)$ are polynomials having real coefficients, and $K = \frac{b_m}{a_n}$ is a constant. The roots of the polynomial $N(s)$ are called *Zeros* and that of $D(s)$ are called *Poles* of the transfer function $H(s)$.

Definition (Bode magnitude plot)**.** *In this plot magnitude of $H(s)$, expressed in the unit of decibels* (**dB**) *i.e.* $20\log_{10}|H|$*, is plotted against frequency having a logarithmic scale.*

6.5 Weak stationarity of stochastic processes

In this section we justify our assumptions of stationarity for the time series considered in Chapter 3. Before proceeding further, we consider definitions of a few statistical measures required to explain weak stationarity in this context.

Definitions

We consider a general stochastic process, i.e. a time-parametrized random variable $\{X(t), t \geq 0\}$, and define the CV and the correlation coefficients (ρ and r_s) as follows.

Definition 1 (Coefficient of variation)**.** *Let $\mathbf{E}[X(t)]$ and $\mathbf{V}[X(t)]$ denote the mean and the variance of $X(t)$. The coefficient of variation $\mathbf{C}[X(t)]$ is defined as*

$$\mathbf{C}[X(t)] = \frac{\sqrt{\mathbf{V}[X(t)]}}{\mathbf{E}[X(t)]} \tag{6.5.1}$$

At steady state, $\mathbf{E}[X(t)] = \mu$ and $\sqrt{\mathbf{V}[X(t)]} = \sigma$ are constants over time and the CV is given by

$$c_v^{ss} = \frac{\sigma}{\mu}. \tag{6.5.2}$$

In this study μ and σ are estimated via Monte Carlo integration with an ensemble of 1000 SSA realizations.

Definition 2 (Correlation coefficient)**.** *The correlation coefficient ρ between two scalar stochastic processes $X(t)$ and $Y(t)$ is defined as the cross-covariance of $X(t)$ and $Y(t)$ normalized to the product of their standard deviations,*

$$\rho = \frac{\mathbf{E}[X(t)Y(t)] - \mathbf{E}[X(t)]\mathbf{E}[Y(t)]}{\sqrt{\mathbf{V}[X(t)]}\sqrt{\mathbf{V}[Y(t)]}}. \tag{6.5.3}$$

If both processes are stationary, $\mathbf{E}[X(t)]$, $\mathbf{E}[Y(t)]$, $\sqrt{\mathbf{V}[X(t)]}$ and $\sqrt{\mathbf{V}[Y(t)]}$ are constants over time. Denoting these constants with μ_X, μ_Y, σ_X, and σ_Y, Equation 6.5.4 reads

$$\rho = \frac{\mathbf{E}[X(t)Y(t)] - \mu_X\mu_Y}{\sigma_X\sigma_Y}. \tag{6.5.4}$$

In this case, $\mathbf{E}[X(t)Y(t)]$ is also constant over time and hence ρ is a time-independent measure, the Pearson correlation coefficient (Pearson, 1896).

Definition 3 (Spearman's correlation coefficient)**.** *The Spearman correlation coefficient, denoted by r_s in this text, is the Pearson's correlation coefficient but for the ranked variables (Spearman, 1987).*

r_s is typically used to describe monotonous but non-linear relations between the variables X and Y.

Definition 4 (Autocovariance & Autocorrelation functions)**.** *The autocovariance of $X(t)$ is defined as the cross-covariance of $X(t)$ with itself; hence, takes the following form:*

$$\Gamma[X(t), X(t+h)] = \mathbf{E}[X(t)X(t+h)] - \mathbf{E}[X(t)]\mathbf{E}[X(t+h)]. \tag{6.5.5}$$

Furthermore, the autocorrelation function is defined as

$$\gamma[X(t), X(t+h)] = \frac{\mathbf{E}[X(t)X(t+h)] - \mathbf{E}[X(t)]\mathbf{E}[X(t+h)]}{\sigma_{X(t)}\sigma_{X(t+h)}}. \tag{6.5.6}$$

Now, coming back to the topic of stationarity in stochastic processes, in Chapter 3, we particularly focus on *weak stationarity* of the time series, meaning that the mean and the autocovariance Γ (or the autocorrelation γ) are time invariant and only depend on the time lag. As a representative example, we argue that the time series of ppX^t for the model $\mathrm{B}^{\mathrm{bio}}_{k^{\mathrm{X}}=k^{\mathrm{Y}}}$ with the same parameter settings as used for Figure 3.10(b), satisfies the conditions for weak stationarity Figure 6.1(a) justifies the stationarity of the mean of ppX^t through visual inspection. Furthermore, Figure 6.1(b) indicates the dependency of γ on the time lag h. As the lag increases, the value of γ decreases, as expected. Finally, Figures 6.1(c) and 6.1(d) depict the time invariance of the value of γ for different time lags.

6.6 Gillespie's stochastic simulation algorithm

The *stochastic simulation algorithm* (SSA) due to Gillespie (Gillespie, 1977) is a Monte Carlo based approach which produces sample time courses from the chemical master equation [2] (Ge and Qian, 2013; Gillespie, 1992). Primarily the algorithm generates two quantities:

[2] The chemical master equation, abbreviated as CME, is a differential-difference equation which is continuous in time and discrete in state space. The state space is defined by the population count of the species in that system. Mathematically, solving CME means solving $\nabla_t P(\mathbf{x}, t) = A_s P(\mathbf{x}, t)$ which

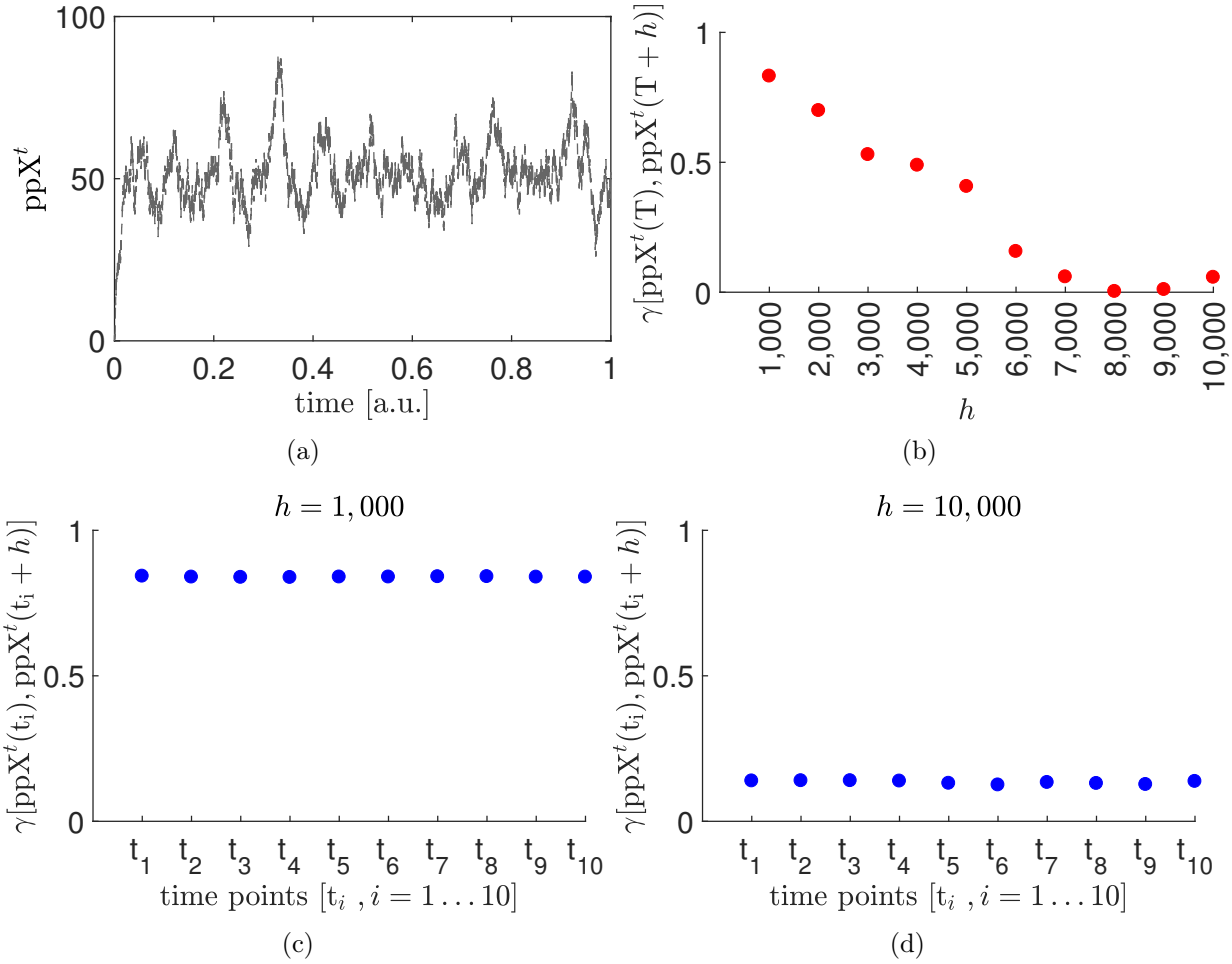


Fig. 6.5.1. Weak stationarity of a sample path of ppX^t for model $\mathrm{B}^{\mathrm{bio}}_{k^{\mathrm{X}}=k^{\mathrm{Y}}}$. (a) Representative sample path of ppX^t for model $\mathrm{B}^{\mathrm{bio}}_{k^{\mathrm{X}}=k^{\mathrm{Y}}}$ shown as in Figure 3.10(b) but for the normalized time range $[0, 1]$. (b) Using Monte Carlo integration from different sample paths, γ is plotted as a function of the time lag (h) for a particular time point T indexed by 10 in the range [0.4-1], where the sample path is visually appread to be mean stationary. (c)-(d) Empirical autocorrelations are plotted for different time points for two constant time lags $h = 1,000$ and $10,000$. The x-axis labels are made symbolic in order to justify the generality of the observation. For numerical purposes, we consider 10 time indices $\{t_i\}_{i=1}^{10}$ as [160 161 162 163 164 165 166 167 168 169 170], respectively. (b)-(d) are generated over 50 SSA realizations taking samples within the time interval $[0.4, 1]$. Rest of the parameters are equal to that for Figure 3.10(b). Original source: Paul and Radde (2018, Figure C.11)

index of the next reaction and time until that particular reaction. Imagine in a chemically reacting system (well-stirred[3] so that reactions can be modeled using a Markov Process

has a solution $P(\mathbf{x}, t) = \exp(A_s t)P(\mathbf{x}, 0)$, where $\mathbf{x}$ is a the state vector. Often is the case when the state-reaction matrix A_s(Munsky and Khammash, 2006) is singular and the state space is infinite dimensional. In that case CME is hard to solve analytically as well as numerically (Sunkara, 2009).

[3]The term well-stirred refers to two conditions. The first one is the *spatial homogeneity* that ensures the positions of the reacting molecules are independent and random variables and are uniformly distributed over the entire volume where the reactions are taking place, and the second one is the *Maxwell-Boltzmann*

- a fundamental assumption of the algorithm), there are n reactions. Choosing the i^{th} reaction is equivalent to rolling a n-sided dice where each side is weighted with the *reaction propensity*, a term that tells how likely a reaction will occur per unit time. The reaction propensity of the i^{th} reaction, denoted by a_i, depends on the order of that reaction. For example, for a first order reaction $S \xrightarrow{c}$, $a_i = c \cdot X_S(t)$, where $X_S(t)$ is the number of molecules of species S at time t. For a second order reaction $S + T \xrightarrow{c}$, $a_i = c \cdot X_S(t)X_T(t)$, and for a dimerization $S + S \xrightarrow{c}$, $a_i = \frac{1}{2}c \cdot X_S(t)(X_S(t) - 1)$ (Higham, 2008).

As mentioned before, the algorithm has two components: index of the next reaction and the time of the next reaction. In order to derive them, let us introduce two quantities: $P_0(\eta|t)$ - probability that no reaction will occur in the time interval $[t, t + \eta)$, and $P_i(\eta|t) = a_i\eta$, probability that i^{th} reaction will occur in the time interval $[t, t + \eta)$. Therefore, $P_0(\eta|t) = 1 - \sum_{i=1}^{n} P_i(\eta|t) = 1 - \sum_{i=1}^{n} a_i\eta$. Now assuming a Markov process i.e. an event occuring in $[t, t + \eta]$ is independent of the event in the time interval $[t + \eta, t + \eta + \delta\eta]$, We get,

$$\begin{aligned} P_0(\eta + \delta\eta|t) &= P_0(\eta|t)P_0(\delta\eta|t + \eta) \\ &= P_0(\eta|t)\left(1 - \sum_{i=1}^{n} a_i\delta\eta\right) \end{aligned}$$

Considering $\delta\eta \to 0$ and rearranging the terms to form the differential equation for $P_0(\eta|t)$, we have the following solution,

$$P_0(\eta|t) = \exp(-A\eta), A = \sum_{i=1}^{n} a_i$$

Now, consider the index for the next reaction as j. Then the probabilty of j^{th} reaction occuring in the time interval $[t + \eta, t + \eta + \delta\eta]$ is given by

$$P(\delta\eta, j|\eta + t)\delta\eta = P_0(\eta|t)P_j(\delta\eta|\eta + t) = P_0(\eta|t)a_j\delta\eta \implies P(\delta\eta, j|\eta + t) = a_j \exp(-A\eta).$$

Rewriting the expression for $P(\delta\eta, j|\eta + t)$ as

$$P(\delta\eta, j|\eta + t) = \frac{a_j}{A} \cdot A \exp(-A\eta),$$

we have

- $\frac{a_j}{A}$ as the next reaction index and
- $A \exp(-A\eta)$ as the time until the next reaction as the density function of continuous random variable having an *exponential distribution.*

Formally, Gillespie's stochastic simulation algorithm proceeds in the following way (Higham,

velocity distribution according to which the Cartesian component of a randomly selected molecule is normally distributed with zero mean and $\frac{k_B T}{m}$ variance, where m is the mass of the molecule and T is the temperature [Source: `https://massimostella.files.wordpress.com/2015/03/mathbis.pdf`]

2008). Although SSA produces probabilistically exact samples from the chemical master

Algorithm 1: Gillespie's stochastic simulation algorithm

Input: $X(0)$: Initial state column vector at $t = 0$; Final time T; ν=Stoichiometric matrix having dimension of # species × # of reactions.
Result: $X(T)$: State vector at time T.

while $t<T$ **do**
  Calculate $\{a_i(X_t)\}_{i=1}^{n}$ and A
  Draw $\psi_1, \psi_2 \sim U(0,1)$
  $j = \arg\min_j \frac{\sum_{i=1}^{j} a_i(X_t)}{A} > \psi_1$ [index of the next reaction]
  $\eta = \ln\left(\frac{1}{\psi_2}\right) \cdot \frac{1}{A}$ [$\rho \sim U(a,b) \implies -\frac{1}{\lambda}\exp\left(\frac{\rho-a}{b-a}\right) \sim \exp(\lambda)$; time interval until the next reaction]
  $X(t+\eta) = X(t) + \nu_j$ [update the state vector]
  $t = t + \eta$ [update the time]
end

equation, it simulates all the successive reaction events in the systems. Therefore, often for a large system, the algorithm is often computationally inefficient (Gillespie, 2007).

6.7 Enzyme processivity in multisite protein phosphorylation

There are several mechanistic aspects of multisite protein phosphorylation such as the order in which phosphate molecules are processed, enzyme processivity, competition between different phosphoforms at low and high enzyme concentrations, the effect of dynamic equilibrium between different conformational changes on the phosphorylated form, and finally the effect of compartmentalization due to localization of kinases in different subcelluar compartments (Salazar and Höfer, 2009). Within the scope of this thesis, we introduce briefly the mechanistic details of phosphorylation due to *enzyme processivity* only. Enzyme processivity is based on the number of phosphorylation or dephosphorylation events taking place during a single encounter between substrate and enzyme. If the number of events is at most one then the mechanism is *distributive* and if it is two or more then the mechanism is *processive* (Salazar and Höfer, 2009).

Distributive mechanism

In a distributive mechanism, at most one modification (phosphorylation or dephosphorylation) takes place at a time. Before the next modification, the enzyme and the substrate

have to be apart or dissociate from each other. Figure 6.7.1 illustrates the mechanism through a schematic representation.

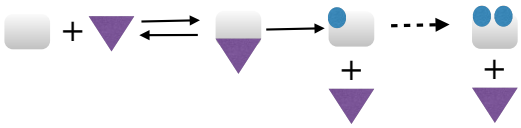


Fig. 6.7.1. Schematic of distributive kinectics. Enzyme is denoted by triangle, substrate and phosphatase molecules are denoted by reactangle and circle, respectively. In distributive mechanism, at most one binding event takes place per enzyme-substrate interaction. The enzyme kinase has to be dissociated form the substrate in order to catalyze the next phosphorylation reaction. The same applies for phosphatase as well.

In the MAPK pathway of the *Xenopus* oocyte, the distributive mechanism has been found to be responsible for the ultrasensitive response of the doubly phosphorylated substrate with respect to the kinase concentration (Burack and Sturgill, 1997; Ferrell and Bhatt, 1997; Huang and Ferrell, 1996). Zhao and Zhang (2001) have demonstrated the same mechanism for the dephosphorylation event mediated by the phosphatase MKP3.

Processive mechanism

Contrary to the distributive mechanism, the processive mechanism involves two or more modification events while the enzyme is bound to the substrate. Figure 6.7.2 illustrates the mechanism through a schematic representation.

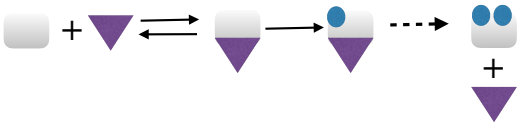


Fig. 6.7.2. Schematic representation of processive kinectics. Phosphorylation at multiple sites of the target protein can take place while the enzyme kinase is bound to the substrate. After all sites of the target protein are phosphorylated, the enzyme kinase dissociates.

An example of a processive mechanism includes phosphorylation of the alternative splicing factor/pre-mRNA-splicing factor (ASF/SF2) protein by Serine/threonine-protein kinase 1 (SRPK1) and dual specificity protein kinase C1k/Sty (Aubol et al., 2003; Velazquez-Dones et al., 2005). ASF/SF2 is a sequence specific splicing factor involved in pre-mRNA splicing (Patwardhan and Miller, 2007). Interestingly, it has been found that for mammalian cells the ERK MAP kinase exhibits a processive dynamics (Aoki et al., 2011) rather than a distributive one as found in the *Xenopus* oocyte. The authors (Aoki et al., 2011) have identified molecular crowding responsible for converting a distributive mechanism to a

processive one resulting in a graded response. For an extensive review on the processive mechanism and its implications on biological systems, interested readers are referred to Patwardhan and Miller (2007).

6.8 Velocity of RNAPII elongation

In this section, we made an attempt to incorporate the velocity of RNAPII to our model following the paper Xu et al. (2015). The authors of the paper assume that nascent mRNAs are elongating with a constant speed denoted by V_{EL}. Therefore the time for elongation or time to produce a complete transcript after initiation is $T_{\mathrm{EL}} = \frac{1}{V_{\mathrm{EL}}}$. The authors provide the expression for the number of nascent mRNA molecules (N_{m}) at a given time of observation t_{obs} as:

$$N_{\mathrm{m}} = \int_{t_{\mathrm{obs}}-T_{\mathrm{EL}}}^{t_{\mathrm{obs}}} n(t)g(t)dt, \tag{6.8.1}$$

where $g(t) = \frac{t}{T_{\mathrm{EL}}}$: contribution function for a single transcript initiated at time t. $n(t) = \sum_i \delta(t_i)$: the total number of transcription events occur within the interval $[t_{\mathrm{obs}} - T_{\mathrm{EL}}\ \ t_{\mathrm{obs}}]$ Now, in our modelling approach, we can realize the velocity of elongation by incorporating the states corresponding to each interaction site designated for polymerase elongation. For example,

$$\mathrm{G}(\dots e_k, \dots)$$

can be rewritten as:

$$\mathrm{G}(\dots e_k\{n_1, \dots, n_N\}[\cdot]),$$

where N is the total number of nucleotides in the coding region. If the velocity (V_{EL}) is given by M nucleotides·time^{-1}, then the elongation time will be $T_{\mathrm{EL}} = \frac{N}{M}$. Now, delay in the elongation process can be realized in the following way (configuration at a particular observation time t):

$$\mathrm{G}(\dots e_k\{n_i\}[p]\ \ \dots, e_m\{n_j\}[q]\dots); i \neq j,\ 1 \leq i,\ j \leq N,$$

where $p, q \in \mathbb{Z}_{>0}$. $e_k\{n_i\}[p]$ denotes that i nucleotides are already been transcribed and $N - i$ nucleotides are still left for transcription by the k^{th} polymerase. Here the contribution function for k^{th} polymerase can be calculated in the following way:

$$g(t) = \frac{i}{N}$$

Bibliography

Ahsendorf, T., Wong, F., Eils, R., and Gunawardena, J. (2014). A framework for modelling gene regulation which accommodates non-equilibrium mechanisms. *BMC Biol*, 12(1):102.

Alderson, D. L. and Doyle, J. C. (2010). Contrasting views of complexity and their implications for network-centric infrastructures. *IEEE Trans Syst Man Cybern A Syst Hum*, 40(4):839–852.

Allen, B. L. and Taatjes, D. J. (2015). The mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol*, 16(3):155.

Angeli, D., Ferrell, J., and Sontag, E. (2004). Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. *Proc Natl Acad Sci U S A*, 101(7):1822–1827.

Aoki, K., Takahashi, K., Kaizu, K., and Matsuda, M. (2013). A quantitative model of ERK MAP kinase phosphorylation in crowded media. *Nat Sci Rep*, 3:1541.

Aoki, K., Yamada, M., Kunida, K., Yasuda, S., and Matsuda, M. (2011). Processive phosphorylation of ERK MAP kinase in mammalian cells. *Proc Natl Acad Sci*, 108(31):12675–12680.

Ardito, F., Giuliani, M., Perrone, D., Troiano, G., and Lo Muzio, L. (2017). The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy. *Int J Mol Med*, 40(2):271–280.

Arkun, Y. and Yasemi, M. (2018). Dynamics and control of the ERK signaling pathway: Sensitivity, bistability, and oscillations. *PloS one*, 13(4):e0195513.

Aubol, B. E., Chakrabarti, S., Ngo, J., Shaffer, J., Nolen, B., Fu, X.-D., Ghosh, G., and Adams, J. A. (2003). Processive phosphorylation of alternative splicing factor/splicing factor 2. *Proc Natl Acad Sci*, 100(22):12601–12606.

Barrett, J. (1963). The use of functionals in the analysis of non-linear physical systems. *Int J Electron*, 15(6):567–615.

Bartman, C. R., Hsu, S. C., Hsiung, C. C.-S., Raj, A., and Blobel, G. A. (2016). Enhancer regulation of transcriptional bursting parameters revealed by forced chromatin looping. *Mol Cell*, 62(2):237–247.

Batchelor, E. and Goulian, M. (2003). Robustness and the cycle of phosphorylation and dephosphorylation in a two-component regulatory system. *Proc Natl Acad Sci U S A*, 100(2):691–96.

Billings, S. and Zhang, H. (1994). Analysing non-linear systems in the frequency domain–ii. the phase response. *Mech Syst Signal Process*, 8(1):45–62.

Birtwistle, M., Rauch, J., Kiyatkin, A., Aksamitiene, E., Dobrzynski, M., Hoek, J., Kolch, W., Ogunnaike, B., and Kholodenko, B. (2012). Emergence of bimodal cell population responses from the interplay between analog single-cell signaling and protein expression noise. *BMC Syst Biol*, 6(109):1–12.

Blüthgen, N. and Legewie, S. (2013). Robustness of signal transduction pathways. *Cell Mol Life Sci*, 70:2259–69.

Boettiger, A. N. and Levine, M. (2009). Synchronous and stochastic patterns of gene activation in the Drosophila embryo. *Science*, 325(5939):471–473.

Boutillier, P., Maasha, M., Li, X., Medina-Abarca, H. F., Krivine, J., Feret, J., Cristescu, I., Forbes, A. G., and Fontana, W. (2018). The kappa platform for rule-based modeling. *Bioinformatics*, 34(13):i583–i592.

Brightman, F. and Fell, D. (2000). Differential feedback regulation of the MAPK cascade underlies the quantitative differences in EGF and NGF signalling in PC12 cells. *FEBS Lett*, 482:169–74.

Burack, W. R. and Sturgill, T. W. (1997). The activating dual phosphorylation of MAPK by MEK is nonprocessive. *Biochemistry*, 36(20):5929–5933.

Caicedo-Casso, A., Kang, H.-W., Lim, S., and Hong, C. (2015). Robustness and period sensitivity analysis of minimal models for biochemical oscillators. *Nat Sci Rep*, 5(13161):1–13.

Cardelli, L., Csikász-Nagy, A., Dalchau, N., Tribastone, M., and Tschaikowski, M. (2016). Noise reduction in complex biological switches. *Nat Sci Rep*, 6(20214):1–11.

Chickarmane, V., Kholodenko, B., and Sauro, H. (2007). Oscillatory dynamics arising from competitive inhibition and multisite phosphorylation. *J Theor Biol*, 244:68–76.

Cho, E. J. and Kim, J. S. (2012). Crowding effects on the formation and maintenance of nuclear bodies: insights from molecular-dynamics simulations of simple spherical model particles. *Biophys J*, 103(3):424–433.

Chubb, J. R., Trcek, T., Shenoy, S. M., and Singer, R. H. (2006). Transcriptional pulsing of a developmental gene. *Curr Biol*, 16(10):1018–1025.

Cohen, P. (2001). The role of protein phosphorylation in human health and disease. *Eur J Biochem*, 268(19):5001–5010.

Cohen, P. (2002). The origins of protein phosphorylation. *Nat Cell Biol*, 4(5):E127–E130.

Conaway, R. C. and Conaway, J. W. (2013). The mediator complex and transcription elongation. *Biochim Biophys Acta Gene Regul Mech*, 1829(1):69–75.

Dangkulwanich, M., Ishibashi, T., Bintu, L., and Bustamante, C. (2014). Molecular mechanisms of transcription through single-molecule experiments. *Chem Rev*, 114(6):3203–3223.

Danos, V., Feret, J., Fontana, W., Harmer, R., and Krivine, J. (2007). Rule-based modelling of cellular signalling. In *International conference on concurrency theory*, pages 17–41. Springer.

Danos, V. and Laneve, C. (2004). Formal molecular biology. *Theor Comput Sci*, 325(1):69–110.

Del Vecchio, D., Ninfa, A. J., and Sontag, E. D. (2008). Modular cell biology: retroactivity and insulation. *Mol Syst Biol*, 4(1):161.

Del Vecchio, D. and Sontag, E. D. (2009). Engineering principles in bio-molecular systems: From retroactivity to modularity. *Eur J Control*, 15(3-4):389–397.

den Breems, N. Y., Nguyen, L. K., and Kulasiri, D. (2014). Integrated signaling pathway and gene expression regulatory model to dissect dynamics of escherichia coli challenged mammary epithelial cells. *BioSystems*, 126:27–40.

Dexter, J. and Gunawardena, J. (2012). Dimerization and bifunctionality confer robustness to the isocitrate dehydrogenase regulatory system in *Escherichia coli*. *J Biol Chem*, 288(8):5770–78.

Dexter, J., Xu, P., Gunawardena, J., and McClean, M. (2015). Robust network structure of the Sln1-Ypd1-Ssk1 three-component phospho-relay prevents unintended activation of the HOG MAPK pathway in *Saccharomyces cerevisiae*. *BMC Syst Biol*, 9(17):1–15.

Dey, S. S., Foley, J. E., Limsirichai, P., Schaffer, D. V., and Arkin, A. P. (2015). Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. *Mol Syst Biol*, 11(5):806.

Dhananjaneyulu, V., Kumar, G., Viswanathan, G. A., et al. (2012). Noise propagation in two-step series MAPK cascade. *PloS one*, 7(5):e35958.

Dieci, G., Bosio, M. C., Fermi, B., and Ferrari, R. (2013). Transcription reinitiation by RNA polymerase iii. *Biochim Biophys Acta (BBA)- Gene Regul Mech*, 1829(3-4):331–341.

Duchi, D., Bauer, D. L., Fernandez, L., Evans, G., Robb, N., Hwang, L. C., Gryte, K., Tomescu, A., Zawadzki, P., Morichaud, Z., et al. (2016). RNA polymerase pausing during initial transcription. *Mol Cell*, 63(6):939–950.

Estrada, J., Wong, F., DePace, A., and Gunawardena, J. (2016). Information integration and energy expenditure in gene regulation. *Cell*, 166(1):234–244.

Feret, J., Danos, V., Krivine, J., Harmer, R., and Fontana, W. (2009). Internal coarse-graining of molecular systems. *Proc Nat Acad Sci*, 106(16):6453–6458.

Ferrell, J. E. and Bhatt, R. R. (1997). Mechanistic studies of the dual phosphorylation of mitogen-activated protein kinase. *J Biol Chem*, 272(30):19008–19016.

Fitz, V., Shin, J., Ehrlich, C., Farnung, L., Cramer, P., Zaburdaev, V., and Grill, S. W. (2016). Nucleosomal arrangement affects single-molecule transcription dynamics. *Proc Natl Acad Sci U.S.A*, 113(45):12733–12738.

Frank, S. (2003). Genetic variation of polygenic characters and the evolution of genetic degeneracy. *Journal of evolutionary biology*, 16(1):138–142.

Frank, S. A. (2007). Maladaptation and the paradox of robustness in evolution. *PLoS One*, 2(10):e1021.

Frank, S. A. (2013). Evolution of robustness and cellular stochasticity of gene expression. *PLoS Biol*, 11(6):e1001578.

Friedlander, T., Mayo, A., Tlusty, T., and Alon, U. (2015). Evolution of bow-tie architectures in biology. *Plos Comput Biol*, 11(3):e1004055.

Friedman, L. J. and Gelles, J. (2012). Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation. *Cell*, 148(4):679–689.

Fritsche-Guenther, R., Witzel, F., Sieber, A., Herr, R., Schmidt, N., Braun, S., Brummer, T., Sers, C., and Blüthgen, N. (2011). Strong negative feedback from Erk to Raf confers robustness to MAPK signalling. *Mol Syst Biol*, 7(489).

Fujioka, A., Terai, K., Itoh, R. E., Aoki, K., Nakamura, T., Kuroda, S., Nishida, E., and Matsuda, M. (2006). Dynamics of the Ras/ERK MAPK cascade as monitored by fluorescent probes. *J Biol Chem*, 281(13):8917–8926.

Garfield, D. A., Runcie, D. E., Babbitt, C. C., Haygood, R., Nielsen, W. J., and Wray, G. A. (2013). The impact of gene expression variation on the robustness and evolvability of a developmental gene regulatory network. *PLoS Biol*, 11(10):e1001696.

Ge, H. and Qian, H. (2013). Chemical master equation. *Encyclopedia of Systems Biology*, pages 396–399.

Ghavi-Helm, Y., Klein, F. A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E. E. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, 512(7512):96.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J Phys Chem*, 81(25):2340–2361.

Gillespie, D. T. (1992). A rigorous derivation of the chemical master equation. *Physica A*, 188(1):404–425.

Gillespie, D. T. (2007). Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem*, 58:35–55.

Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036.

Gomez-Uribe, C., Verghese, G. C., and Mirny, L. A. (2007). Operating regimes of signaling cycles: statics, dynamics, and noise filtering. *PLoS Comput Biol*, 3(12):e246.

Grieco, L., Calzone, L., Bernard-Pierrot, I., Radvanyi, F., Kahn-Perlés, B., and Thieffry, D. (2013). Integrative modelling of the influence of MAPK network on cancer cell fate decision. *PLoS Comput Biol*, 9(10).

Gunawardena, J. (2005). Multisite protein phosphorylation makes a good threshold but can be a poor switch. *Proc Natl Acad Sci U S A*, 102(41):14617–22.

Hahn, S. (2004). Structure and mechanism of the RNA polymerase ii transcription machinery. *Nat Struct Mol Biol*, 11(5):394.

Hammerstein, P., Hagen, E. H., Herz, A. V., and Herzel, H. (2006). Robustness: A key to evolutionary design. *Biol Theory*, 1(1):90–93.

Hart, Y., Madar, D., Yuan, J., Bren, A., Mayo, A., Rabinowitz, J., and Alon, U. (2011). Robust control of nitrogen assimilation by a bifunctional enzyme in *E. Coli*. *Mol Cell*, 41:117–27.

Hersen, P., McClean, M., Mahadevan, L., and Ramanathan, S. (2008). Signal processing by the HOG MAP kinase pathway. *Proc Natl Acad Sci U S A*, 105(20):7165–70.

Hespanha, J. P. (2018). *Linear systems theory*. Princeton university press.

Higham, D. J. (2008). Modeling and simulating chemical reactions. *SIAM review*, 50(2):347–368.

Hu, D. and Yuan, J.-M. (2006). Time-dependent sensitivity analysis of biological networks: Coupled MAPK and PI3K signal transduction pathways. *J Phys Chem*, 110:5361–70.

Huang, C.-Y. and Ferrell, J. E. (1996). Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc Natl Acad Sci*, 93(19):10078–10083.

Jen, E. (2005). *Robust design: a repertoire of biological, ecological, and engineering case studies*. Oxford University Press.

Johnson, K. A. and Goody, R. S. (2011). The original Michaelis constant: translation of the 1913 Michaelis–Menten paper. *Biochemistry*, 50(39):8264–8269.

Kholodenko, B. (2000). Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase casacde. *Eur J Biochem*, 267:1583–1588.

Kielkiewicz, M. (1970). Cascade connection of non-linear systems. *Int J Control*, 11(6):1005–1010.

Kim, J. K. and Marioni, J. C. (2013). Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol*, 14(1):R7.

Kim, Y., Coppey, M., Grossman, R., Ajuria, L., Jiménez, G., Paroush, Z., and Shvartsman, S. Y. (2010). MAPK substrate competition integrates patterning signals in the Drosophila embryo. *Curr Biol*, 20(5):446–451.

Kim, Y., Paroush, Z., Nairz, K., Hafen, E., G., J., and Shvartsman, S. (2011a). Substrate-dependent control of MAPK phosphorylation in vivo. *Mol Syst Biol*, 467(7).

Kim, Y., Paroush, Z., Nairz, K., Hafen, E., Jiménez, G., and Shvartsman, S. Y. (2011b). Substrate-dependent control of MAPK phosphorylation in vivo. *Mol Syst Biol*, 7(1):467.

Kirch, J., Thomaseth, C., Jensch, A., and Radde, N. E. (2016). The effect of model rescaling and normalization on sensitivity analysis on an example of a MAPK pathway model. *EPJ Nonlinear Biomed Phys*, 4(1):3.

Kitano, H. (2007). Towards a theory of biological robustness. *Mol Syst Biol*, 3(1):137.

Klein, P., Pawson, T., and Tyers, M. (2003). Mathematical modeling suggests cooperative interactions between a disordered polyvalent ligand and a single receptor site. *Curr Biol*, 13(19):1669–1678.

Klipp, E. and Liebermeister, W. (2006). Mathematical modeling of intracellular signaling pathways. *BMC Neurosci*, 7(1):S10.

Klumpp, S. and Hwa, T. (2008). Stochasticity and traffic jams in the transcription of ribosomal RNA: Intriguing role of termination and antitermination. *Proc Nat Acad Sci*, 105(47):18159–18164.

Kolch, W. (2005). Coordinating ERK/MAPK signalling through scaffolds and inhibitors. *Nat Rev Mol Cell Biol*, 6:827–838.

Kolch, W., Calder, M., and Gilbert, D. (2005). When kinases meet mathematics: the systems biology of MAPK signalling. *FEBS Lett*, 579:1891–95.

Kornberg, R. D. (2007). The molecular basis of eukaryotic transcription. *Proc Natl Acad Sci*, 104(32):12955–12961.

Krauss, G. (2006). *Biochemistry of signal transduction and regulation*. John Wiley & Sons.

Krivine, J., Danos, V., and Benecke, A. (2009). Modelling epigenetic information maintenance: A kappa tutorial. In *International Conference on Computer Aided Verification*, pages 17–32. Springer.

Kumar, N., Singh, A., and Kulkarni, R. V. (2015). Transcriptional bursting in gene expression: analytical results for general stochastic models. *PLoS Comput Biol*, 11(10):e1004292.

Kwon, S. (2013). Single-molecule fluorescence in situ hybridization: quantitative imaging of single RNA molecules. *BMB reports*, 46(2):65.

Lee, T.-H. and Maheshri, N. (2012). A regulatory role for repeated decoy transcription factor binding sites in target gene expression. *Mol Syst Biol*, 8(1):576.

Legewie, S., Schoeberl, B., Blüthgen, N., and Herzel, H. (2007). Competing docking interactions can bring about bistability in the MAPK cascade. *Biophys J*, 93:2279–88.

Lerner, E., Ingargiola, A., Lee, J. J., Borukhov, S., Michalet, X., and Weiss, S. (2017). Different types of pausing modes during transcription initiation. *Transcription*, 8(4):242–253.

Li, L., Zhao, G.-D., Shi, Z., Qi, L.-L., Zhou, L.-Y., and Fu, Z.-X. (2016). The Ras/Raf/MEK/ERK signaling pathway and its role in the occurrence and development of hcc. *Oncol Lett*, 12(5):3045–3050.

Liu, X., Kraus, W. L., and Bai, X. (2015). Ready, pause, go: regulation of RNA polymerase ii pausing and release by cellular signaling pathways. *Trends Biochem Sci*, 40(9):516–525.

MacNeil, L. and Walhout, A. J. (2011). Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res*, pages gr–097378.

Malleshaiah, M. K., Shahrezaei, V., Swain, P. S., and Michnick, S. W. (2010). The scaffold protein ste5 directly controls a switch-like mating decision in yeast. *Nature*, 465(7294):101.

Markevich, N., Hoek, J., and Kholodenko, B. (2004). Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *J Cell Biol*, 164(3):353–59.

MATLAB (2016). *version 9.1.0 (R2016b)*. The MathWorks Inc., Natick, Massachusetts.

Meng, H. and Bartholomew, B. (2018). Emerging roles of transcriptional enhancers in chromatin looping and promoter-proximal pausing of RNA polymerase ii. *J Biol Chem*, 293(36):13786–13794.

Mettetal, J., Muzzey, D., Gomez-Uribe, C., and van Oudenaarden, A. (2008). The frequency dependence of osmo-adaptation in *Saccharomyces cerevisiae*. *Science*, 319(5862):482–84.

Michaelis, L. and Menten, M. (1913). Die kinetik der invertinwirkung Biochem z 49: 333–369.

Millikan, R. A. and Bishop, E. S. (1917). *Elements of electricity: a practical discussion of the fundamental laws and phenomena of electricity and their practical applications in the business and industrial world*. American Technical Society.

Milo, R. (2013). What is the total number of protein molecules per cell volume? a call to rethink some published values. *Bioessays*, 35(12):1050–1055.

Mugler, A., Walczak, A. M., and Wiggins, C. H. (2009). Spectral solutions to stochastic models of gene expression with bursts and regulation. *Phys Rev E*, 80(4):041921.

Munsky, B. and Khammash, M. (2006). The finite state projection algorithm for the solution of the chemical master equation. *J Chem Phys*, 124(4):044–104.

Munsky, B., Neuert, G., and Van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187.

Myers, L. C. and Kornberg, R. D. (2000). Mediator of transcriptional regulation. *Annu Rev Biochem*, 69(1):729–749.

Mylona, A., Theillet, F.-X., Foster, C., Cheng, T. M., Miralles, F., Bates, P. A., Selenko, P., and Treisman, R. (2016). Opposing effects of Elk-1 multisite phosphorylation shape its response to ERK activation. *Science*, 354(6309):233–237.

Nam, K. M. (2018). Computing steady-state mRNA copy-number distributions. unpublished.

Nicolas, D., Phillips, N. E., and Naef, F. (2017). What shapes eukaryotic transcriptional bursting? *Mol BioSyst*, 13(7):1280–1290.

Nicolas, D., Zoller, B., Suter, D. M., and Naef, F. (2018). Modulation of transcriptional burst frequency by histone acetylation. *Proc Nat Acad Sci*, page 201722330.

O'Shaughnessy, E., Palani, S., Collins, J., and Sarkar, C. (2011). Tunable signal processing in synthetic MAP kinase cascades. *Cell*, 144(1):119–31.

Ouldridge, T. and ten Wolde, P. (2014). The robustness of proofreading to crowding-induced pseudo-processivity in the MAPK pathway. *Biophys J*, 107:2425–35.

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., and Van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. *Nat Genet*, 31(1):69.

Padovan-Merhar, O., Nair, G. P., Biaesch, A. G., Mayer, A., Scarfone, S., Foley, S. W., Wu, A. R., Churchman, L. S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and dna copy number through independent global transcriptional mechanisms. *Mol Cell*, 58(2):339–352.

Patwardhan, P. and Miller, W. T. (2007). Processive phosphorylation: mechanism and biological importance. *Cell Signal*, 19(11):2218–2226.

Paul, D., Dehkordi, L. K. F., von Scheven, M., Bischoff, M., and Radde, N. (2016). Structural design with biological methods: optimality, multi-functionality and robustness. In *Biomimetic Research for Architecture and Building Construction*, pages 341–360. Springer.

Paul, D. and Radde, N. (2016). Robustness and filtering properties of ubiquitous signaling network motifs. *IFAC-PapersOnLine*, 49(26):120–127.

Paul, D. and Radde, N. (2018). The role of stochastic sequestration dynamics for intrinsic noise filtering in signaling network motifs. *J Theor Biol*, 455:86 – 96.

Payne, J. L. and Wagner, A. (2015). Mechanisms of mutational robustness in transcriptional regulation. *Front Genet*, 6:322.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philos Trans R Soc Lond A, containing papers of a mathematical or physical character*, 187:253–318.

Peccoud, J. and Ycart, B. (1995). Markovian modeling of gene-product synthesis. *Theor Popul Biol*, 48(2):222–234.

Prabakaran, S., Lippens, G., Steen, H., and Gunawardena, J. (2012). Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding. *Wiley Interdiscip Rev Syst Biol Med*, 4(6):565–583.

Qiao, L., Nachbar, R., Kevrekidis, I., and Shvartsman, S. (2007). Bistability and oscillations in the Huang-Ferrell model of MAPK signaling. *PLoS Comput Biol*, 3(9):1819–26.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biol*, 4(10):e309.

Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226.

Ramsey, S., Orrell, D., and Bolouri, H. (2005). Dizzy: stochastic simulation of large-scale genetic regulatory networks. *J Bioinform Comput Biol*, 3(02):415–436.

Roberts, J. W. (2014). Molecular basis of transcription pausing. *Science*, 344(6189):1226–1227.

Roignant, J.-Y. and Treisman, J. E. (2009). Pattern formation in the Drosophila eye disc. *Int J Dev Biol*, 53(5-6):795.

Roskoski Jr, R. (2010). Raf protein-serine/threonine kinases: structure and regulation. *Biochem Biophys Res Commun*, 399(3):313–317.

Salazar, C. and Höfer, T. (2009). Multisite protein phosphorylation–from molecular mechanisms to kinetic models. *Febs J*, 276(12):3177–3198.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons.

Samoilov, M., Plyasunov, S., and Arkin, A. (2005). Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proc Natl Acad Sci U S A*, 102:2310–15.

Santos, S., Verveer, P., and Bastiaens, P. (2007). Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nat Cell Biol*, 9(3).

Schwabe, A., Rybakova, K. N., and Bruggeman, F. J. (2012). Transcription stochasticity of complex gene regulation models. *Biophy J*, 103(6):1152–1161.

Senecal, A., Munsky, B., Proux, F., Ly, N., Braye, F. E., Zimmer, C., Mueller, F., and Darzacq, X. (2014). Transcription factors modulate c-fos transcriptional bursts. *Cell Rep*, 8(1):75–83.

Shah, R. and Vecchio, D. D. (2017). Signaling architectures that transmit unidirectional information despite retroactivity. *Biophys J*, 113(3):728 – 742.

Shahrezaei, V. and Swain, P. S. (2008). Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci*.

Shinar, G. and Feinberg, M. (2010). Structural sources of robustness in biochemical reaction networks. *Science*, 327:1389–91.

Shinar, G. and Feinberg, M. (2011). Design principles for robust biochemical reactiion networks: What works, what cannot work, and what might almost work. *Math Biosci*, 231:39–48.

Shinar, G., Rabinowitz, J., and Alon, U. (2009). Robustness in glyoxylate bypass regulation. *PLoS Comput Biol*, 5(3):e1000297.

Spearman, C. (1987). The proof and measurement of association between two things. *The Am J Psychol*, 100(3/4):441–471.

Sunkara, V. (2009). The chemical master equation with respect to reaction counts. In *Proc. 18th World IMACS/MODSIM Congress*, pages 703–707. Citeseer.

Suter, D. M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., and Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474.

Varma, A., Morbidelli, M., and Wu, H. (2005). *Parametric sensitivity in chemical systems*. Cambridge University Press.

Velazquez-Dones, A., Hagopian, J. C., Ma, C.-T., Zhong, X.-Y., Zhou, H., Ghosh, G., Fu, X.-D., and Adams, J. A. (2005). Mass spectrometric and kinetic analysis of ASF/SF2 phosphorylation by SRPK1 and Clk/Sty. *J Biol Chem*, 280(50):41761–41768.

Ventura, A. C., Jackson, T. L., and Merajver, S. D. (2009). On the role of cell signaling models in cancer research. *Cancer Res*, 69.

Ventura, A. C., Jiang, P., Van Wassenhove, L., Del Vecchio, D., Merajver, S. D., and Ninfa, A. J. (2010). Signaling properties of a covalent modification cycle are altered by a downstream target. *Proc Natl Acad Sci USA*, 107.

Ventura, A. C., Sepulchre, J. A., and Merajver, S. D. (2008). A hidden feedback in signaling cascades is revealed. *PLoS Comput Biol*, 4.

Ventura, J.-J. and Nebreda, Á. R. (2006). Protein kinases and phosphatases as therapeutic targets in cancer. *Clin Transl Oncol*, 8(3):153–160.

Volterra, V. (1930). *Theory of functionals*. Blackie.

Whitmarsh, A. J. (2007). Regulation of gene transcription by mitogen-activated protein kinase signaling pathways. *Biochim Biophys Acta Mol Cell Res*, 1773(8):1285–1298.

Whitmarsh, A. J. and Davis, R. J. (2016). Multisite phosphorylation by MAPK. *Science*, 354(6309):179–180.

Williams, R. L., Lawrence, D. A., et al. (2007). *Linear state-space control systems*. John Wiley & Sons.

Wynn, M. L., Ventura, A. C., Sepulchre, J. A., García, H. J., and Merajver, S. D. (2011). Kinase inhibitors can produce off-target effects and activate linked pathways by retroactivity. *BMC Syst Biol*, 5(1):156.

Xiong, W. and Ferrell, J. (2003). A positive-feedback-based bistable memory module that governs cell fate decision. *Nature*, 426:460–65.

Xu, H., Sepúlveda, L. A., Figard, L., Sokac, A. M., and Golding, I. (2015). Combining protein and mRNA quantification to decipher transcriptional regulation. *Nat Methods*, 12(8):739.

Xu, H., Skinner, S. O., Sokac, A. M., and Golding, I. (2016). Stochastic kinetics of nascent RNA. *Phys Rev Lett*, 117(12):128101.

Yang, S.-H., Sharrocks, A. D., and Whitmarsh, A. J. (2003). Transcriptional regulation by the MAP kinase signaling cascades. *Gene*, 320:3–21.

Zhang, H. and Billings, S. (1993). Analysing non-linear systems in the frequency domain–i. the transfer function. *Mech Syst Signal Process*, 7(6):531–550.

Zhang, J., Nie, Q., He, M., and Zhou, T. (2013). An effective method for computing the noise in biochemical networks. *J Chem Phys*, 138(8):02B615.

Zhao, Y. and Zhang, Z.-Y. (2001). The mechanism of dephosphorylation of extracellular signal-regulated kinase 2 by mitogen-activated protein kinase phosphatase 3. *J Biol Chem*, 276(34):32382–32391.

Zhou, T. (2013). *Encyclopedia of Systems Biology*, chapter Oscillation Amplitude, pages 1616–1616. Springer, New York, NY.