

Mathematical Modelling and Parameter Inference of Genetic Regulatory Networks

A thesis submitted for the degree of
Doctor of Philosophy

by

Qianqian Wu

School of Mathematical Sciences
Monash University
Australia

September 2015

Contents

Copyright notice	ix
Abstract	xi
List of publications	xv
Acknowledgment	xix
1 Introduction	3
1.1 Background	3
1.2 Genetic Regulatory Networks	5
1.3 Mathematical Modelling	7
1.3.1 Boolean networks	7
1.3.2 Bayesian/ Graphical networks	8
1.3.3 Deterministic models	8
1.3.4 Stochastic models	9
1.4 Parameter Inference	15
1.4.1 Bayesian inference method	17
1.4.2 Approximate Bayesian methods	18
1.4.3 Parameter identification	23
1.5 Objectives	24
1.6 Thesis Outline	26
2 A Two-variable Model for Stochastic Modelling of Chemical Events with Multi-step Reactions	35

2.1	Introduction	35
2.2	New Modelling Method for Multi-step Reaction System	39
2.3	Ordinary Differential Equation Model	41
2.4	Application to mRNA Degradation	45
2.5	Conclusion	52
3	Stochastic Modelling of Biochemical Systems of Multi-step Reactions using a Simplified Two-Variable Model	59
3.1	Introduction	59
3.2	Results and Discussion	63
3.2.1	A new two-variable model	63
3.2.2	Determination of probability function	67
3.2.3	mRNA decay dynamics: a case study for gene <i>RPL30</i>	71
3.3	Conclusion	77
3.4	Methods	79
3.4.1	Simulation algorithm for the probability function	79
3.4.2	Ordinary differential equation model	81
3.4.3	An algorithm for simulating systems including two-variable model	82
4	Stochastic Modelling of Regulatory Networks using State-dependent Time Delay	89
4.1	Introduction	89
4.2	Methods	92
4.2.1	Multi-step chemical reaction system	92
4.2.2	State-dependent time delay	93
4.2.3	SSA with state-dependent time delay	94
4.3	Results	96
4.3.1	State-dependent time delay	96

4.3.2	Formula for calculating time delay	97
4.3.3	Time delay of mRNA degradation	99
4.3.4	Time delay in gene expression	101
4.4	Discussion and Conclusion	108
4.5	Supplementary Information	109
4.5.1	Multi-step chemical reaction system	109
4.5.2	Algorithm for calculating time delay	111
4.5.3	Formulation of time delay	113
5	Approximate Bayesian Computation for Estimating Rate Constants in Biochemical Reaction Systems	123
5.1	Introduction	123
5.2	Method	126
5.3	Results	129
5.3.1	Decay-dimerization model	129
5.3.2	Prokaryotic auto-regulatory gene network	134
5.4	Conclusion	139
6	Approximate Bayesian Computation Schemes for Parameter Infer- ence of Discrete Stochastic Models using Simulated Likelihood Den- sity	145
6.1	Introduction	145
6.2	Results and discussion	148
6.2.1	The first test system with four reactions	148
6.2.2	The second test system with eight reactions	152
6.3	Conclusion	159
6.4	Methods	160
6.4.1	ABC SMC algorithm	160
6.4.2	ABC using simulated likelihood density	162

7	Sensitivity and Robustness Analysis for Stochastic Model of Nanog Gene Regulatory Network	171
7.1	Introduction	171
7.2	Method	175
7.2.1	Mathematical model	176
7.2.2	Framework for sensitivity and robustness analysis	178
7.3	Results	181
7.3.1	Deterministic behaviour	181
7.3.2	Stochastic behaviour	182
7.3.3	Sensitivity analysis	184
7.3.4	Robustness analysis	187
7.4	Discussion and Conclusion	191
8	An Integrated Approach to Infer Dynamic Protein-gene Interactions: A Case Study of the Human P53 Protein	197
8.1	Introduction	197
8.2	Materials and Methods	200
8.2.1	Experimental dataset	200
8.2.2	Top-down approach	201
8.2.3	Bottom-up approach	203
8.3	Results	211
8.3.1	Inference of a network of eight genes with full connections	211
8.3.2	Inference of the network of eight genes using predicted network structure from top-down approach	212
8.3.3	Inference of the network of eight interactions with extended regulations	213
8.3.4	Inference of the network of eight regulations with auto-regulation	217
8.3.5	Network structure perturbation - edge deletion	218

8.3.6	A Comparison study to an earlier inference method	220
8.3.7	Inference of a medium network of 21 genes	223
8.4	Discussions	224
8.5	Supplementary Information	228
8.6	Conclusion	237
9	Conclusion	241
9.1	Contributions of the Thesis	241
9.2	Future Directions	244
	Bibliography	277

Copyright notice

©The author (2015). Except as provided in the Copyright Act 1968, this thesis may not be reproduced in any form without the written permission of the author.

Abstract

Mathematical modelling opens the door to a rich pathway to study the dynamic properties of biological systems. Among the many biological systems that would benefit from mathematical modelling, improving our understanding of gene regulatory networks has received much attention from the fields of computational biology and bioinformatics. To understand system dynamics of biological networks, mathematical models need to be constructed and studied. In spite of the efforts that have been given to explore regulatory mechanisms among gene networks, accurate description of chemical events with multi-step chemical reactions still remains a challenge in biochemistry and biophysics. This dissertation is aimed at developing several novel methods for describing dynamics of multi-step chemical reaction systems. The main idea is introduced by a new concept for the location of molecules in the multi-step reactions, which is used as an additional indicator of system dynamics. Additionally, novel idea in the stochastic simulation algorithm is used to calculate time delay exactly, which shows that the value of time delay depends on the system states. All of these innovations alter the focus of originally complex multi-step structures towards defining novel simplified structures, which simplifies the modelling process significantly. Research results yield substantially more accurate results than published methods.

Apart from the well-established knowledge for modelling techniques, there are still significant challenges in understanding the dynamics of systems biology. One

of the major challenges in systems biology is how to infer unknown parameters in mathematical models based on experimental datasets, in particular, when data are sparse and networks are stochastic. To tackle this challenge, parameters estimation techniques using Approximate Bayesian Computation (ABC) for chemical reaction system and inference method for dynamic network have been investigated. This dissertation discusses developed ABC methods that have been tested on two stochastic systems. Results on artificial data show certain promising approximations for the unknown parameters in the systems. While unknown parameters are difficult and sometimes even impossible to measure with biological experiments, instead we can study the influence of parameter variation on system properties. Robustness and sensitivity are two major measurements to describe the dynamic properties of a system against the variation of model parameters. For stochastic models of discrete chemical reaction systems, although these two properties have been studied separately, no work has been done so far to investigate these two properties together. In this dissertation, An integrated framework has been proposed to study these two properties for the Nanog gene network simultaneously. It successfully identifies key coefficients that have more impacts on the network dynamics than the others.

The proposed inference method to infer dynamic protein-gene interactions is applied to a case study of the human P53 protein, which is a well-known biological network for cancer study. Investigating the dynamics for such regulatory networks through high throughput experimental data has become more popular. To tackle the hindrances with large number of unknown parameters when building detailed mathematical models, a new integrated method is proposed by combining a top-down approach using probability graphical models and a bottom-up approach using differential equation models. Model simulation error, Akaike's information criterion, parameter identifiability and robustness properties are used as criteria to select the optimal network. Results based on random permutations of input

gene network structures provide accurate prediction and robustness property. In addition, a comparison study suggests that the proposed approach has better simulation accuracy and robustness property than the earlier one. In particular, the computational cost is significantly reduced. Overall, the new integrated method is a promising approach for investigating the dynamics of genetic regulations.

List of publications

1. Wu Q, Smith-Miles K, Tian T. 2012. A two-variable model for stochastic modelling of chemical events with multi-step reactions. In: *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*. pp. 1–6, doi: 10.1109/BIBM.2012.6392681
2. Wu Q, Smith-Miles K, Zhou T, Tian T. 2013b. Stochastic modelling of biochemical systems of multi-step reactions using a simplified two-variable model. *BMC Systems Biology* 7(4): S14, doi: 10.1186/1752-0509-7-S4-S14, URL <http://dx.doi.org/10.1186/1752-0509-7-S4-S14>
3. Wu Q, Tian T. 2015. Stochastic modelling of regulatory networks using state-dependent time delay (to be submitted)
4. Wu Q, Smith-Miles K, Tian T. 2013a. Approximate bayesian computation for estimating rate constants in biochemical reaction systems. In: *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*. pp. 416–421, doi: 10.1109/BIBM.2013.6732528
5. Wu Q, Smith-Miles K, Tian T. 2014. Approximate bayesian computation schemes for parameter inference of discrete stochastic models using simulated likelihood density. *BMC bioinformatics* 15(S12): S3, doi: 10.1186/1471-2105-15-S12-S3, URL <http://www.biomedcentral.com/1471-2105/15/S12/S3>

6. Wu Q, Jiang F, Tian T. 2015. Sensitivity and robustness analysis for stochastic model of nanog gene regulatory network. *International Journal of Bifurcation and Chaos* 25(07): 1540 009, doi: 10.1142/S021812741540009X, URL <http://www.worldscientific.com/doi/abs/10.1142/S021812741540009X>
7. Wang J, Wu Q, Tian T. 2015. An integrated approach to infer dynamic protein-gene interactions: a case study of the human p53 protein (submitted for publication)

PART A: General Declaration

Monash University

Declaration for thesis based or partially based on conjointly published or unpublished work

General Declaration

In accordance with Monash University Doctorate Regulation 17.2 Doctor of Philosophy and Research Master's regulations the following declarations are made:

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes 4 original papers published in peer reviewed journals/conference proceedings and 3 unpublished publications. The core theme of the thesis is mathematical modelling and parameter inference. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the candidate, working within the School of Mathematical Sciences, Monash University, under the supervision of A/Prof. Tianhai Tian and Prof. Kate Smith-Miles.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.

In the case of Chapters 2-8 my contribution to the work involved the following:

Thesis chapter	Publication title	Publication status*	Nature and extent of candidate's contribution
2	A two-variable model for stochastic modelling of chemical events with multi-step reactions	Published	Developed, established and verified the method Wrote programming codes and the article
3	Stochastic modelling of biochemical systems of multi-step reactions using a simplified two-variable model	Published	Developed, established and verified the method Wrote programming codes and the article
4	Stochastic modelling of regulatory networks using state-dependent time delay	To be Submitted	Developed, established and verified the method Wrote programming codes and the article
5	Approximate bayesian computation for estimating rate constants in biochemical reaction systems	Published	Developed, established and verified the method Wrote programming codes and the article
6	Approximate bayesian computation schemes for parameter inference of discrete stochastic models using simulated likelihood density	Published	Developed, established and verified the method Wrote programming codes and the article
7	Sensitivity and robustness analysis for stochastic model of nanog gene regulatory network	In press	Developed, established and verified the method Wrote programming codes and the article
8	An integrated approach to infer dynamic protein-gene interactions: a case study of the human p53 protein	Submitted	Verified the method, wrote part of the programming codes and the article

I have / have not (circle that which applies) renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Signed: 

Date: 29/04/2015

Acknowledgment

I would never have been able to finish my dissertation without the guidance of my supervisors, help from colleagues and friends, and support from my family.

Foremost, I would like to express my sincere gratitude to my supervisors Prof. Tianhai Tian and Prof. Kate Smith-Miles for their continuous support of my Ph.D study and research, for their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Also, It is my pleasure to acknowledge all the colleagues I have met and worked with in the School of Mathematical Sciences at Monash. Having experienced so many years studying there, they have lead me to the wonderful world of Math, helped me to develop my background in math and further fulfilled my early research career.

Many friends have helped me stay sane through these difficult years overseas. Their support and care helped me overcome setbacks and stay focused on my study. I greatly value their friendship and I deeply appreciate their belief in me.

Most importantly, none of this would have been possible without the love and patience of my family. My immediate family, to whom this dissertation is dedicated to, has been a constant source of love, concern, support and strength all these

years. I would like to express my heart-felt gratitude to my parents who have aided and encouraged me throughout this endeavour.

Though only my name appears on the cover of this dissertation, a great many people have contributed to its production. I owe my gratitude to all of you who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

List of Figures

1.1	Example of a genetic regulatory system, consisting of a network of three genes a , b , and c , repressor proteins A , B , C , and D , and their mutual interactions.	6
2.1	Simulation results from the ODE model: (A) for the optimal q value against different X , (B) averaged optimal values for q against n . . .	44
2.2	The SSA simulation results with $s_1 = k_1 = 1$: (A, C) three simulations of X and L values over t for the detailed multi-step reaction model, (B, D) three simulations of X and L values over t for approximated model with two variables, (E) the mean value of X with 10,000 simulations for both models, (F) the mean value of L with 10,000 simulations for both models.	49
2.3	The SSA simulation results with $s_1 = k_1 = 0$: (A, C) three simulations of X and L values over t for the detailed multi-step reaction model, (B, D) three simulations of X and L values over t for approximated model with two variables, (E) the mean value of X with 10,000 simulations for both models, (F) the mean value of L with 10,000 simulations for both models.	51
3.1	The probability for the firing of the last reaction ($B_n \xrightarrow{k_n} P$).	66
3.2	Simulated exact probabilities and two approximated probabilities for the firing of the last reaction.	68

3.3	Simulation results from probability approach: (A) the optimal values of q with different n ; (B) the optimal values of q with different X ; (C) the averaged optimal values of q against n	69
3.4	Relationship between n and q : Dashed-line: estimated relationship from stochastic simulations; dash-dot-line: relationship derived from the ODE model; Solid-line: $q = 0.5n$	71
3.5	Simulated mRNA degradation dynamics using the estimated model parameters: (A) Deterministic simulations for mRNA numbers from the <i>ACT1</i> construct (green dash-line: the one-step model ($k = 0.0276$), red solid-line: the two-variable model with the optimal initial length ($k = 0.112, L = 371$), black dot-line: the two-variable model with the averaged initial length $L = \frac{nX}{2}$, blue dots: experimental data); (B) Deterministic simulations for mRNA numbers from the <i>RPL30</i> construct (green dash-line: the one-step model ($k = 0.0343$), red solid-line: the two-variable model with the optimal initial length ($k = 0.167, L = 473$), black dot-line: the two-variable model with the averaged initial length $L = \frac{nX}{2}$ ($k = 0.161$), blue dots: experimental data); (C) Stochastic simulations of the two-variable model for the <i>ACT1</i> construct (red dot-line: initial $X_0 = 5, k = 0.115, L = 19$, black dash-dot-line: initial $X_0 = 10, k = 0.111, L = 37$, blue dots: experimental data); (D) Stochastic simulations of the two-variable model for the <i>RPL30</i> construct (red dot-line: initial $X_0 = 5, k = 0.171, L = 24$, black dash-dot-line: initial $X_0 = 10, k = 0.166, L = 47$, blue dots: experimental data).	73

3.6	mRNA degradation dynamics of gene <i>RPL30</i> construct <i>ACT1</i> in single cells: (A,C) three simulations of X and L values over t for the detailed multi-step reaction model. (B,D) three simulations of X and L values over t for the two-variable model. For the detailed multi-step model, rate constants are $s_1 = 0, s_2 = \dots = s_{10} = 0.112$, and initial molecular numbers are $([A],[B],[BC1],\dots,[BC7]) = [4,3,3,5,15,20,20,15,15]$. For the two-variable model, rate constant is $k = 0.112$, initial conditions $(X_0,L_0) = (100,371)$	78
4.1	Calculated time delay using stochastic simulations of the multi-step reactions process (4.2.1): (A) Time delay for the decay of each molecule based on different initial number x_{10} but null initial imaginary species y_0 . Index i means the delay of the i -th molecule. (B) Time delay for the decay of the first molecule based on different values of x_{10} and y_0 . (Solid-line: $x_{10} = 5$, dash-line: $x_{10} = 10$, dash-dot-line: $x_{10} = 20$, dot-line: $x_{10} = 40$).	98
4.2	Simulation of mRNA degradation for gene <i>RPL30</i> using the state-dependent delay model: (A) Construct <i>ACT1</i> using estimated parameters $k = 0.1260, y_0 = 23, D = 1.7184$. (B) Construct <i>RPL30</i> using estimated parameters $k = 0.1260, y_0 = 17, D = 1.7525$. (Solid line: averaged mRNA numbers based on 1000 simulations; dash-dot line: experimental data assuming $s_0 = 100$).	102
4.3	Distributions of estimated model parameters for gene <i>RPL30</i> degradation: (A, B, C) Construct <i>ACT1</i> . (D, E, F) Construct <i>RPL30</i>	103
4.4	Simulation of gene transcription for gene <i>SWI5</i> using the state-dependent delay model: (A) mRNA copy number in nucleus. (B) mRNA copy number in cytosol. (dot: experimental data; circle: simulations). Estimated parameters are $a = 3.9137, b = 8.2969, \tau_1 = 36.9431, \tau_2 = 2.0682, k_2 = 2247.9, k_3 = 1.0926$	106

4.5	Distributions of estimated model parameters for gene SWI5 transcription: (A) Transcription delay. (B) Degradation rate constant k_3	107
4.6	A new algorithm for calculating time delay that is dependent on system state: (A) Estimated optimal values of C_2 based on different system states (x_{10}, y_0) that match time delay showing in Figure 4.1 (B). Each line represents the optimal value of C_2 for a particular value of x_{10} . For a fixed value of y_0 , the smaller the value of x_{10} is, the smaller the value of C_2 becomes. (Solid-line: $x_{10} = 5$, dash-line: $x_{10} = 10$, dash-dot line: $x_{10} = 20$, dot-line: $x_{10} = 50$). (B) Values of α . (blue-solid line: estimated values based on simulated time delay in Figure 4.1 (B); red-dash line: prediction from $\alpha = 3.25 + 7.5/x_1$). (C) value of β . (blue-solid line: estimated values based on simulated time delay in Figure 4.1 (B); red-dash line: prediction from $\beta = 11.8 + 8.2x_1$). (D) The difference between the predicted values of C_2 and optimal values of C_2 in Figure 4.6 (A). (Solid-line: $x_{10} = 5$, dash-line: $x_{10} = 10$, dash-dot line: $x_{10} = 20$, dot-line: $x_{10} = 50$). . .	117
4.7	Simulation of mRNA degradation for gene RPL30 using different initial mRNA number s_0 . For each construct, parameters k and D are the same as those in Figure 4.2. The value of y_0 is proportional to the value of s_0 : (A, B, C, D) Construct ACT1. (E, F, G, H) Construct RPL30. (A, E) $s_0 = 20$. (B, F) $s_0 = 50$. (C, G) $s_0 = 150$. (D, H) $s_0 = 200$. (Solid-line: simulation. Dash-dot line: experimental data).	118
5.1	Simulated experimental data for system dynamics in a time length of 30 with step size Δt of 1 (Blue star for S_1 , green circle for S_2 , and red cross for S_3).	131
5.2	Probability distributions of estimated rate constant of c_4 over four iterations ((A): Iteration 2; (B): 3; (C): 4; (D): 5).	132

5.3	The averaged error of estimated parameters and mean count number of iterations with step size Δt of 3 ((A), (B)) and 5 ((C), (D)).	133
5.4	Simulated experimental data for system dynamics in a time length of 50 with step size Δt of 1 (Blue star for <i>DNA</i> ; green circle for <i>DNA.P₂</i> and red cross for <i>mRNA</i> black; cyan square for <i>P</i> ; black x-mark for <i>P₂</i>).	137
5.5	Probability distribution of estimated rate constant of c_7 over four iterations ((A):Iteration 2; (B): 3; (C):4; (D):5).	138
5.6	The averaged error of estimated parameters and mean count number of iterations with step size Δt of 1 ((A), (B)) and 5 ((C), (D)).	139
6.1	Simulated experimental data for system dynamics in a time length of 30 with step size Δt of 3: Blue star for S_1 , green circle for S_2 , and red cross for S_3	149
6.2	Probabilistic distributions of estimated rate constant of c_1 over four iterations using algorithm 1. (A): Iteration 2; (B): 3; (C): 4; (D): 5.	150
6.3	Simulated molecular numbers for system 2 in a time length of 50 with step size Δt of 1: (A): DNA numbers; (B): numbers of <i>DNA.P₂</i> ; (C): Red line for the numbers of <i>mRNA</i> black and cyan dash-dotted line for the numbers of <i>P</i> ; (D): numbers of <i>P₂</i>	155
6.4	Probabilistic distributions of the estimated rate constant c_7 over four iterations using algorithm 1. (A):Iteration 2; (B): 3; (C):4; (D):5.	158
7.1	Network diagram for the Nanog gene regulatory network	176
7.2	Bifurcation diagram of the deterministic model (7.3.1) for four parameters. Solid and dash-dot lines are two stable steady states but dash line in the middle is the unstable steady state. (A) parameter s_3 . (B) s_4 . (C) d_N . (D) p	183

7.3	A stochastic simulation of the proposed stochastic model (7.2.1) using the first set of parameters together with $s_4 = 150$ and random variable p defined by (7.3.3).	185
7.4	Density functions for variables [Oct4-Sox2, Nanog, Rex1, FGF4/Erk] with variations in parameter d_E	186
7.5	Derivatives of density functions for variables [Oct4-Sox2, Nanog, Rex1, FGF4/Erk] with parameter d_E	187
7.6	Sensitivity values for variables [Oct4-Sox2, Nanog, Rex1, FGF4/Erk] with all parameters	188
7.7	Sensitivity values of the stochastic model with all parameters. The indexes (1 ~ 11) are for (1: s_{12} , 2: s_3 , 3: s_4 , 4: s_5 , 5: s_6 , 6: s_7 , 7: d_{OS} , 8: d_N , 9: d_R , 10: d_E and 11: p_0)	189
7.8	Robustness analysis showing the percentages of time points when the network maintains a low expression level of gene Rex1 using different values of a particular parameter. (A) Parameters d_{OS} and s_5 . (B) D_{NN} , p_{00} and s_7 . (C) d_{EE} and s_4 . (D) s_3	190
7.9	Simulations of Nanog number using different model parameters. (A) $s_3 = 0.075$. (B) $s_3 = 0.125$. (C) $s_4 = 30$. (D) $s_4 = 50$	191
8.1	Simulations of the gene network of eight genes: <i>RAD21</i> , <i>pcnA</i> , <i>RAD23B</i> , <i>DDB2</i> (dash-star: microarray data; solid-line: simulation of the fully connected model; dash-line: simulation of the core network predicted by the GGF).	205

8.2	The graphic models of eight genes using different significant levels: Gene-gene interaction networks are predicted by applying GGF on 8 genes that are related to DNA repair pathway: (A) there are 8 regulations among 8 genes, significance level $p < 0.05$; (B) there are 17 regulations among 8 genes, significance level $p < 0.09$, in which includes 8 regulations from the (A). Number on each edge is partial correlation coefficient between the two genes. The network is visualized by Cytoscape software.	214
8.3	A gene-gene interaction network was predicted by applying GGF on 21 genes that related to DNA repair pathway: There are 26 regulations among the genes, significance level $p < 0.009$. Number on each edge is partial correlation coefficient between the two genes. The network is visualized by Cytoscape software.	225
8.4	Simulations of the gene network with 21 genes (dash-star: microarray data; solid-line: simulation of the fully connected model; dash-line: simulation of the core network predicted by the GGF).	226
8.5	A workflow chart of Gaussian Graphical Model with Forward search algorithm (GGF).	229
8.6	Simulations of the gene network model of eight genes: The dynamics of four genes were presented in the paper. Here are the remaining four genes: PTTG1, XPC, RAD51C, RPS27L (dash-star: microarray data; solid-line: simulation of the fully connected model; dash-line: simulation of the core network predicted by the GGF with eight mutual regulations).	230

8.7	Simulations of the gene network of 21 genes: The dynamics of four genes of this network was presented in the paper in Figure (8.4). Here are the other 9 genes. (dash-star: microarray data; solid-line: simulation of the fully connected model; dash-line: simulation of the core network predicted by the GGF).	231
8.8	The graphic model of eight genes using a published inference method: Gene-gene interaction network was predicted by using the inference method in (Äijö and Lähdesmäki, 2009). The inferred network is a full matrix whose diagonal elements are zero. To match the inferred network in Fig. (8.2A), we selected the top 16 edges that have the largest values of the posterior probabilities.	232
8.9	Simulation of the gene network model in Fig. (8.7): (Dash-star: microarray data; red-dash-line: simulation of the network model).	233
8.10	Simulation of the gene network of eight genes by merging the two networks in Fig. (8.2A) and Fig. (8.8) together: This merged network has 11 edges. (dash-star: microarray data; red-dash-line: simulation of the merged gene network).	234

List of Tables

2.1	Reactions and kinetic rates of the simplified stochastic model. The rate constants s_i are in the unit of 1/sec.	46
3.1	Reactions, kinetic rates and propensity functions of the simplified stochastic model. The rate constants s_i are in the unit of 1/sec. . .	74
3.2	Estimated parameters for the stochastic model of <i>RPL30</i> and <i>ACT1</i> mRNA degradations (Ratio= L_0/nX).	76
5.1	Comparison of averaged error for estimated rate constants over five iterations with different simulation numbers and step sizes	134
5.2	Comparison of mean count number for estimated rate constants over five iterations with different simulation numbers and step sizes	135
6.1	Comparison of averaged error and mean count number for estimated rate constants over five iterations using algorithms 1 and 2 with simulation number of 10 for system 1. Tests are experimented under different strategies of discrepancy tolerance such as $\alpha = 0.1, 0.05$ or varies over iterations (AE:Averaged Error; MN: Mean count Number).	153
6.2	Comparison of averaged error and mean count number for estimated rate constants of system 2 using algorithms 1 and 2. Three strategies are used to choose the discrepancy tolerance α : a fixed value of $\alpha = 0.05$; varying α values; and $\alpha = \epsilon_k$ (denoted as same ϵ_k); varying α values that are smaller than ϵ_k (denoted as diff. ϵ_k). (AE:Averaged Error; MN: Mean count Number).	157

8.1 AIC and robustness property of mathematical model of eight genes: Numerical results are presented as the AIC to experimental data, average behaviour of robustness property and standard deviation (STD) of robustness property, which are based on the average of five sets of estimated parameters. (G_i, G_j) means the network by adding the mutual regulation of gene i and gene j to the core network ($p < 0.05$). "Plus 5 regulations" is the network by adding 5 edges, namely (G_4, G_5) , (G_1, G_4) , (G_3, G_7) , (G_5, G_7) , (G_1, G_2) . (G_1 : RAD21, G_2 : pcnA, G_3 : RAD23B, G_4 : DDB2, G_5 : PTTG1, G_6 : XPC, G_7 : RAD51C, G_8 : Rps271 215

8.2 AIC and robustness property of the network ($p < 0.05$) by adding an auto-positive/negative regulation to the core network: $G_i (+)$: add auto-positive regulation of gene i to the core network; $(-)$: add auto-negative regulation. RBN: robustness. 218

8.3 AIC and robustness property of the reduced network with 8 genes by removing one mutual regulation from the core model in Fig. (8.2A): (G_i, G_j) means the network by removing mutual regulation between gene i and gene j from the core network. 220

8.4 AIC and robustness property of the reduced network with 8 genes by removing one one-way regulation from the core model in Fig. (8.1A): $(G_i \leftarrow G_j)$ represent the removing of the regulation from gene j to gene i , namely by letting $a_{ij} = b_{ij} = 0$. RBN: robustness. 221

8.5 Simulation error and robustness property of mathematical models of eight genes: (Network 1: the gene network inferred from the network structure in Fig. (8.2A). Network 2: the gene network inferred from the network structure in Fig. (8.7); Network 3: the gene network inferred from the network structure that is merged from the networks in Fig. (8.2A) and Fig. (8.7)). 222

8.6	Model parameters of the fully connected network with 8 genes: Coefficient a_{ij}, b_{ij} and time delay τ_i in (8.2.3); Degradation rate d_i ; Basal transcription rate c_i ; Initial condition $x_i(0)$	235
8.7	Model parameters of the core network with 8 genes in Fig. (8.1A): Coefficient a_{ij}, b_{ij} and time delay τ_i in (8.2.3); Degradation rate d_i ; Basal transcription rate c_i ; Initial condition $x_i(0)$	236

Chapter 1
Introduction

Chapter 1

Introduction

1.1 Background

Biological systems change over time. The concept of this change includes both as the system oscillates or otherwise moves in some consistent behaviour, which may also change from one time instant to another (Small, 2012). As biological system exhibit rich dynamic behaviour over a large range of time and space scales, the study for such complex systems in a unified framework has been recognized recently as a new scientific discipline. To study the complex systems with exponentially growth of biological data, universal simplifications are particularly important and as well as to integrate and organize the data into coherent descriptive models (Bar-Yam, 1997; Peleg *et al.*, 2005).

Once I have read through the famous parable of six blind men inspecting an elephant in the book (Haefner, 2005), which tells as follows:

“They are asked to identify the object before them which they cannot see. One man, feeling the elephant’s leg, thinks he is touching a tree trunk. Another, grasping the elephant’s trunk, thinks he is holding a snake. A third, standing near the moving ear, thinks it is a

large, feathered fan. And so it goes for the other men touching the tusk, the side, and the tail of the elephant. Each man gave a different description of the same object, but none was correct.”

From this story, we can see how each man are creating a model of the system they are observing while non of them are fully correct but provide an approximation. In the words of George Box: “ *All models are wrong, some models are useful.*”. A good model should be simple enough to be useful, but not so simple that it no longer reflects useful information of reality (Small, 2012). In system biology, such a good model can shed insight into complex biological processes and suggest new directions for research. The ability to predict system dynamical behaviour with a model helps evaluate model completeness as well as developing our understanding of the mechanisms of biological processes. One of the major challenges currently facing modern biology is to build a systematic understanding of biological networks based on the established foundation of molecular characterization of cell components. Mathematical models and computer simulations, which are powerful and predictive tools, offer insight into the dynamics of temporal and spatial biological systems such as genetic regulatory networks, cell signalling pathways and metabolic pathways.

Building models from data in this study is considered to be a two-part problem, namely constructing the model structure and inference of parameters inside the model. The following sections will review through an extensive tour of one of the most important biological systems - “genetic regulatory networks”, and introduce mathematical methods for simulating such networks as well as parameter estimation methods.

1.2 Genetic Regulatory Networks

With only few exceptions, all cells in an organism share the same genetic material. Genome has been regarded as a dominant position to control cellular processes such as cell divisions and replications. Understanding how genes are expressed and regulated in such processes has been of high interest for the last decades, therefore, we need to study deeply from sequences of nucleotides coding for proteins to regulatory systems that determines the gene expressions (De Jong, 2002). The expression of a gene are activated and inactivated by random association and dissociation events (Paulsson, 2005; Lockhart and Winzeler, 2000). These stochastic fluctuations in gene expression lead to considerable differences in the level of expression between genetically identical cells (Kaern *et al.*, 2005), which play crucial roles in biological processes (Heitzler and Simpson, 1991). Several studies have measured variability in protein and messenger RNA levels, and discovered strong connections between noise and gene regulation mechanisms. Due to the vast range of gene activities, gene regulation is of high complexity. Transcription is universally the first step toward expressing a gene, which is a highly regulated process and understanding this transcription regulation is of fundamental importance. For protein-coding genes, post-transcriptional steps, including pre-mRNA processing, mRNA transportation and translation, also play significant roles in regulating gene expression (Ma, 2010). It's been clearly stated that the regulation of gene expression is achieved through genetic regulatory networks (GRNs) of interactions between DNA, RNA, proteins and small molecules, which leads to great attention for GRN over the last few years (Iba and Mimura, 2002). Due to the study of Human Genome project, genetic information has become increasingly available, which makes studying biological system from genetic aspects achievable.

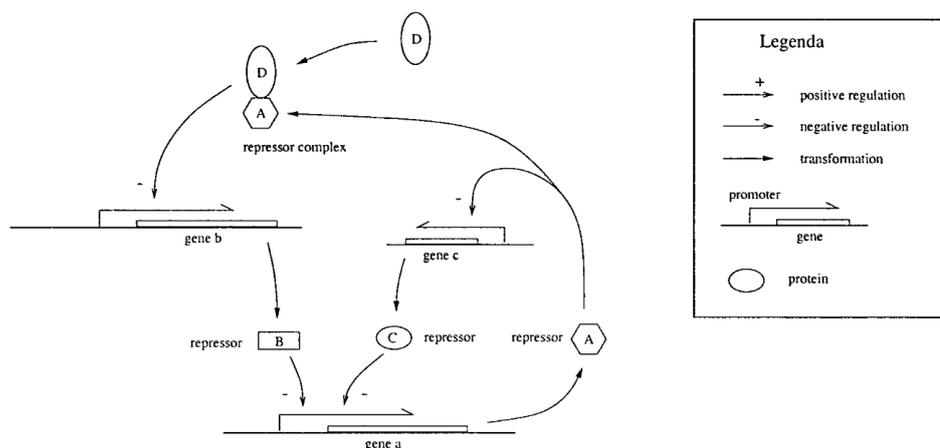


Figure 1.1: Example of a genetic regulatory system, consisting of a network of three genes *a*, *b*, and *c*, repressor proteins *A*, *B*, *C*, and *D*, and their mutual interactions.

Gene regulatory networks (GRNs) explain the interconnections between genomic entities that govern the regulation of gene expression. A simple example for a regulatory network, involving three genes that code for proteins inhibiting the expression of other genes, is shown in Fig. 1.1 (De Jong, 2002). We can find that proteins *B* and *C* independently repress gene *a* while proteins *A* and *D* interact to form a substance that binds to a regulatory site of gene *b*. This figure proposes a basic form for a GRN. More complex graphical conventions to represent cellular networks can be found in Kohn (1999, 2001). Since biological regulatory networks are extremely detailed with numerous interactions, a single mathematical model to represent the whole biological regulatory system is generally not feasible. The focus of the modelling can be capturing interactions between RNA expressions, protein-protein interactions, or interactions between metabolites. Usually, only parts of the regulome (genes, proteins, and metabolites involved in gene regulation) such as transcription factors, enhancers, and microRNA are made explicit in a mathematical model of a GRN (Pal *et al.*, 2012). The modelling can be deterministic capturing the average behaviour of a colony of cells or stochastic capturing the inherent noise in biological systems. Furthermore, the models can be fine scale

or coarse-scale (Karlebach and Shamir, 2008). Next section presents a brief review of various techniques for modelling GRNs.

1.3 Mathematical Modelling

“All models are wrong, but some are useful”. - George E. P. Box

A large number of approaches have been proposed to model the behaviour of GRNs (De Jong, 2002; Szallasi *et al.*, 2006; Cai and Wang, 2007). All modelling methodologies have strengths and weaknesses regarding their ease and fidelity of capturing biological system dynamics. Typically, these techniques can be broadly classified into continuous and discrete modelling strategies based on how the solution space is acquired. Additional classification into deterministic and stochastic models is an alternative method that divides systems based on whether they contain a degree of “randomness” that allows for multiple solutions to the same initial conditions (Walpole *et al.*, 2013). Various computational models have been developed for regulatory network analysis and the most commonly used models are reviewed as follows.

1.3.1 Boolean networks

The most basic and simplest discrete modelling methodology was introduced by Kauffman and Thomas in 1973 (Glass and Kauffman, 1973; Thomas, 1973). It allows to rely on purely qualitative data and can be analysed using a broad range of well-established mathematical methods. In such a discrete Boolean network, each node x_i can attain two alternative states: on (1) or off (0), which forms as 0-1 vectors describing the system’s state/global state. For example, a gene can be described as expressed or not expressed at any time, then a node is updated depending on the current states of all nodes in the network: $x_i \leftarrow f(x_1, \dots, x_N)$. Modelling

regulatory networks using Boolean network has become popular and successful to describe yeast cell-cycle etc. (Wittmann *et al.*, 2009; Davidich and Bornholdt, 2008). However, it's unable to count for continuous changes of concentrations or the exact timing of regulatory events (Karlebach and Shamir, 2008). It uses discretised data, which to some extent subjects the formalism to information loss from the data discretization.

1.3.2 Bayesian/ Graphical networks

Probabilistic reasoning based on incomplete prior biological knowledge and current observations has been applied to build models of GRNs, which is commonly known as Bayesian networks or graphical models (Jordan, 1998). This technique was first introduced by Kauffman (Kauffman, 1969) that is based on conditional dependencies between sets of variables (De Jong, 2002). They are defined by a family of conditional distributions and a set of corresponding parameters (Omony, 2014). Dynamic Bayesian network (DBN) as an extension of the Bayesian network incorporates time dynamics into the GRN and allows feedback relations among genes to be modeled (Murphy *et al.*, 1999; Friedman *et al.*, 2000). Applications for such networks can be found extensively in the work of (Kim *et al.*, 2003) and (Zou and Conzen, 2005). However, this approach involves using numerous assumptions and some of which are neither robust nor adequate (Spirtes *et al.*, 2000).

1.3.3 Deterministic models

Rather than discrete-measured experimental data, biological experiments usually produce real and continuous measurements; therefore using real-valued parameters over a continuous timescale is essential while modeling. Nonlinear ordinary differential equations and piecewise linear differential equations have

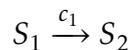
been proposed as continuous fine-scale deterministic models for GRNs. Differential equation (DE) models assume that species concentration vary continuously and deterministically (Pal *et al.*, 2012). Continuous modelling strategies include using systems of ordinary differential equations (ODEs) and partial differential equations (PDEs) to solve for steady state solutions (Fallahi-Sichani *et al.*, 2012b; Scheff *et al.*, 2011; Adra *et al.*, 2010; Greenstein and Winslow, 2011; Fallahi-Sichani *et al.*, 2012a; Quo *et al.*, 2011; Laise *et al.*, 2011). One type of continuous model, namely continuous linear model, is defined as each regulator contributes to the input of the regulation function independently of the other regulators, which do not require extensive knowledge about regulatory mechanisms. There are some other networks where the regulators are transcription factors and the levels of genes are determined by real-valued, non-linear regulation functions that take the Michaelis-Menten form (Klipp *et al.*, 2008). In this case, non-linear models arise.

1.3.4 Stochastic models

Stochastic modelling and simulation of biological processes are problems of high interest today, which is our main focus as well. In real world systems, despite deterministic networks, where the state of the system is determined by the current state and external inputs, stochastic effects play an important role. For example, in yeast, the number of mRNA molecules is close to one copy per cell for some genes (Holstege *et al.*, 1998). This indicates that it's likely that there is a considerable intrinsic noise element present, that is to say some cells have more mRNA molecules of the given species present than others. Therefore, modelling a cell by using continuous concentrations should be modelling an ensemble of cells by the average values of stochastic variables. It has been demonstrated that the stochastic effects are important for the phage λ switch decision between lysis and lysogeny (McAdams and Arkin, 1997). And more experimental studies present the measurements for the level of intrinsic noise in eukaryotic cells (Paulsson, 2004;

Raser and O’Shea, 2004). Simulating a stochastic model is computationally more expensive as the simulations have to run several times to provide good results of the system behaviour. However, stochastic models are the best choices for the systems in which small number of molecules and some random effects are involved (Schlitt and Brazma, 2007).

Stochastic and discrete fine-scale models are commonly known as stochastic master equation (SME) models. To explain an SME model, consider a system with N molecular species $\{S_1, \dots, S_N\}$ and M different reaction channels $\{R_1, \dots, R_M\}$, where the state of the system is defined by $\{\mathbf{X}(t) = X_1(t), \dots, X_N(t)\}$, where $X_i(t)$ is the number of molecules of species S_i in the system at time t . Each reaction channel R_j can be characterized by a propensity function a_j and a state change vector $\mathbf{v}_j = (v_{1j}, \dots, v_{Nj})$, where $a_j(\mathbf{x})dt$ is the probability for one R_j reaction to occur in the next infinitesimal time interval $[t; t + dt)$ given $\mathbf{X}(t) = \mathbf{x}$, v_{ij} is the change in the molecular population S_i induced by one reaction R_j [A rigorous derivation of the chemical master equation]. The value of propensity function depends on the populations of the reactant populations and a reaction probability rate constant c_j , where $c_j dt$ is the probability that a randomly chosen pairs of R_j reactant molecules will react in the next infinitesimal time dt and a_j is the product of c_j and the number of all possible combinations of R_j reactant molecules. For example: For



we have $a_j(\mathbf{x}) = c_1 x_1$, and $\mathbf{v}_j = (-1, 1, 0, \dots, 0)$. The chemical master equation (CME) is obtained once the propensity functions is determined as follows:

$$\frac{\partial P(\mathbf{x}, t | \mathbf{x}_0, t_0)}{\partial t} = \sum_{j=1}^M [a_j(\mathbf{x} - \mathbf{v}_j) P(\mathbf{x} - \mathbf{v}_j, t | \mathbf{x}_0, t_0) - a_j(\mathbf{x}) P(\mathbf{x}, t | \mathbf{x}_0, t_0)]$$

where \mathbf{v} is known as the stoichiometric matrix, $P(\mathbf{x}, t | \mathbf{x}_0, t_0)$ denotes the probability for $\mathbf{X}(t)$ to be \mathbf{x} given that $\mathbf{X}(t_0) = \mathbf{x}_0$. We note that CME is essentially an ODE

whose dimension is given by the number of all possible combinations of states of \mathbf{x} (Cao and Samuels, 2009).

An important question in stochastic modelling is how to develop stochastic models by introducing stochastic processes into deterministic models for the external and/or internal noise. Numerical methods for simulating chemical reaction systems is discussed next. These methods are the theoretical basis for designing stochastic models. The Stochastic Simulation Algorithm (SSA) represents a discrete modelling approach and an essentially exact procedure for numerically simulating the time evolution of a well-stirred reaction system (Gillespie, 1977). The advances in stochastic modelling of genetic regulatory networks and cell signalling transduction pathways have stimulated growing research interests in the development of effective methods for simulating chemical reaction systems. These effective simulation methods in return provide innovative methodologies for designing stochastic models of biological systems.

Stochastic simulation algorithm

The stochastic simulation algorithm (SSA) is a statistically exact procedure for generating the time and index of the next occurring reaction in accordance with the current values of the propensity functions. In each step, two random numbers are generated to determine the time interval and the index of the next reaction. There are several forms of this algorithm. The widely used direct method works in the following manner.

Method 1: the direct method (Gillespie, 1977).

Step 1: Calculate the values of propensity functions $a_j(\mathbf{x})$ based on the system state \mathbf{x} at time t and $a_0(\mathbf{x}) = \sum_{j=1}^M a_j(\mathbf{x})$.

Step 2: Generate a sample r_1 of the uniformly distributed random variable $U(0, 1)$, and determine the time of the next reaction

$$\mu = \frac{1}{a_0(\mathbf{x})} \ln\left(\frac{1}{r_1}\right).$$

Step 3: Generate an independent sample r_2 of $U(0, 1)$ to determine the index k of the next reaction occurring in $[t, t + \mu)$,

$$\sum_{j=1}^{k-1} a_j(\mathbf{x}) < r_2 a_0(\mathbf{x}) \leq \sum_{j=1}^k a_j(\mathbf{x}).$$

Step 4: Update the state of the system by

$$\mathbf{x}(t + \mu) = \mathbf{x}(t) + \nu_k. \tag{1.3.1}$$

Step 5: Go to Step 1 if $t + \mu \leq T$, where T is the end time point. Otherwise, the system state $\mathbf{x}(T) = \mathbf{x}(t)$.

Another exact method is the first reaction method which uses M random numbers at each step to determine the possible reaction time of each reaction channel (Gillespie, 1976). The reaction firing in the next step is that needing the smallest reaction time. Comparing to the direct method, the first reaction method is not efficient since it discards $M - 1$ random numbers at each step. To improve the efficiency of the first reaction method, Gibson and Bruck (2000) proposed the next reaction method by recycling the generated random numbers. The putative step size of a reaction channel is updated based on the step size of this channel at the previous step and values of the propensity function at these two steps. In addition, a so-called dependency graph was designed to reduce the computing time of

propensity functions. Numerical results indicated that the next reaction method is effective for simulating systems with many species and reaction channels.

Stochastic simulation algorithm assumes that the next reaction will fire in the next reaction time interval $[t, t + \mu)$ with small values of μ . For systems including both fast and slow reactions, however, this assumption may not be valid if the slow reactions take a much longer time than the fast reactions. The large reaction time of slow reactions should be realized by time delay τ if we hope to put both fast and slow reactions in a system consistently and to study the impact of slow reactions on the system dynamics (Monk, 2003). Recently, the delay stochastic simulation algorithm (DSSA) was designed to simulate chemical reaction systems with time delays (Barrio *et al.*, 2006; Bratsun *et al.*, 2005; Cai, 2007). These methods have been used to validate stochastic models for biological systems with slow reactions (Roussel and Zhu, 2006; Schlicht and Winkler, 2008). However, compared with the significant progress in designing simulation methods for biological systems without time delay (Gillespie, 2007; Pahle, 2009), only a few simulation methods have been designed to improve the efficiency of the DSSA (Leier *et al.*, 2008; Bayati *et al.*, 2009). Similar to the effective methods for simulating biological systems without time delay, it is expected the progress in designing effective methods for simulating systems with time delay will also provide methodologies for modelling biological systems with time delay. DSSA works in the following manner.

Method 2: the delay method (Barrio *et al.*, 2006).

Step 1: Calculate the values of propensity functions $a_j(\mathbf{x})$ based on the system state \mathbf{x} at time t and $a_0(\mathbf{x}) = \sum_{j=1}^M a_j(\mathbf{x})$.

Step 2: Generate a sample r_1 of the uniformly distributed random variable $\mathbf{U}(0, 1)$, and determine the time of the next reaction

$$\mu = \frac{1}{a_0(\mathbf{x})} \ln\left(\frac{1}{r_1}\right).$$

Step 3: Generate an independent sample r_2 of $\mathbf{U}(0, 1)$ to determine the index k of the next reaction occurring in $[t, t + \mu)$,

$$\sum_{j=1}^{k-1} a_j(\mathbf{x}) < r_2 a_0(\mathbf{x}) \leq \sum_{j=1}^k a_j(\mathbf{x}).$$

Step 4: If delayed reactions are scheduled within $[t, t + \mu)$, then let j be the delayed reaction scheduled next at time $t + \tau$. Update the state of the system by

$$\mathbf{x}(t + \tau) = \mathbf{x}(t) + \nu_j, \tag{1.3.2}$$

and

$$t = t + \tau. \tag{1.3.3}$$

Else if k is not a delayed reaction then update the state of the system by

$$\mathbf{x}(t + \mu) = \mathbf{x}(t) + \nu_k. \tag{1.3.4}$$

Else record time, $t + \mu + \tau$, for delayed reaction k and

$$t = t + \mu. \tag{1.3.5}$$

Step 5: Go to Step 1 if $t + \mu \leq T$, where T is the end time point. Otherwise, the system state $\mathbf{x}(T) = \mathbf{x}(t)$.

1.4 Parameter Inference

The topic of parameter estimation in dynamical systems, which is a wide area involving many different aspects of statistics as well as numerical analysis, at the same time being closely linked to mathematical model building and experimental design. Using ordinary differential equations (ODEs) that describe the evolution over time of certain quantities of interest is quite a popular approach once the pathway structure is given, the corresponding equations are relatively easy to write down using widely accepted kinetic laws, such as the law of mass action or the Michaelis-Menten law (Lillacci and Khammash, 2010). In general the equations depend on several parameters, some of which are reaction rates, and production and decay coefficients with physical meanings, which are seldom known. It is a challenging problem to infer such ODE parameters from gene expression data since the ODEs do not have analytic solutions and the time-course gene expression data are usually sparse and associated with large noise. A better knowledge of kinetic rate constants in the modelling of chemical reactions can help in choosing operating conditions that favour the desired products. Estimating parameters in systems modelled by ODEs is both computationally intensive as well as numerically challenging due to a variety of undesirable characteristics. This is not only the case for modeling with ODEs, often the models contain a number of parameters that cannot be measured directly or calculated by applying established laws of nature, and therefore must be estimated from experimental data.

In the last fifteen years, parameter inference problem has become significantly important in the systems biology community (Lillacci and Khammash, 2010). Inference methods including optimization methods and Bayesian inferences have been approached for estimating unknown parameters. Several optimization techniques, such as linear and nonlinear least-squares fitting (Mendes and Kell, 1998),

simulated annealing (Brooks and Morgan, 1995), genetic algorithms (Srinivas and Patnaik, 1994), and evolutionary computation (Ashyraliyev *et al.*, 2008) are extensively utilized to identify unknown parameters of systems biology models. Global optimization (GO) methods can be roughly classified as deterministic (Horst and Tuy, 1996; Grossmann, 1996; Esposito and Floudas, 2000) and stochastic strategies (Ali *et al.*, 1997; Törn *et al.*, 1999). Stochastic methods for global optimization ultimately rely on probabilistic approaches, which have weak theoretical guarantees of convergence to the global solution. Deterministic methods are those that can provide a level of assurance that the global optimum will be located, and several important advances in the GO of certain types of nonlinear dynamic systems have been made recently (Esposito and Floudas, 2000; Singer *et al.*, 2001; Papamichail and Adjiman, 2002). However, it should be noted that, although deterministic methods can guarantee global optimality for certain GO problems, no algorithm can solve general GO problems with certainty in finite time (Boender and Romeijn, 1995). One of the main problems associated with optimization methods is that they tend to be computationally expensive and may not perform well if the noise in the measurements is significant. Estimation techniques like linear iterative models, stochastic optimization methods and constrained linear and nonlinear regression models are often used in GRN (Dimitrova *et al.*, 2011; Steggles *et al.*, 2007; Almeida and Voit, 2003; Zhan and Yeung, 2011; Singhania *et al.*, 2011; Rodriguez-Fernandez *et al.*, 2006; Chou and Voit, 2009). Each approach has its own strengths and weaknesses many of which are strongly linked to the data quality and modelling approach. Genomic, proteomic and other -omic data types are prone to noise and/or have missing data (Omony, 2014). Most methods focus on small-sized networks because of the computational challenges associated with larger networks. However, the need to accurately describe molecular mechanisms in biochemical systems cannot be understated. To achieve such high performance descriptions, parameters have to be accurately and precisely identified.

1.4.1 Bayesian inference method

Bayesian inference methods have more power to solve problems when modelling biological systems where molecular species are present in low copy numbers and noise exists (Raj and van Oudenaarden, 2008; Wilkinson, 2007). The parameter estimation for stochastic models has been extensively explored in financial mathematics (Johannes and Polson, 2003) and has been applied to biological systems in a frequentist maximum likelihood (Reinker *et al.*, 2006) and Bayesian framework (Golightly and Wilkinson, 2005, 2006; Wilkinson, 2011). Bayesian methods can extract information from noisy or uncertain data, which includes both measurement noise and intrinsic noise (McAdams and Arkin, 1999). In Bayesian statistical inference, a prior probability distribution that benefits from previous knowledge, is the probability distribution that expresses one's uncertainty quantity. The main advantage is its ability to infer the whole probability distributions of the parameters, rather than just a point estimate. Also, they can handle estimation of stochastic systems with no substantial modification to the algorithms (Toni *et al.*, 2009). The main obstacle to their application is computational, since analytical approaches are not feasible for non-trivial problems and numerical solutions are also challenging due to the need to solve high-dimensional integration problems. Nonetheless, the most recent advancements in Bayesian computation, such as Markov chain Monte Carlo techniques (Brooks, 1998), ensemble methods (Brown and Sethna, 2003; Battogtokh *et al.*, 2002), and sequential Monte Carlo methods that do not require likelihoods (Toni *et al.*, 2009; Sisson *et al.*, 2007) have been successfully applied to biological systems. Maximum-likelihood estimation (Müller *et al.*, 2004; Bortz and Nelson, 2006) has also been extensively applied. Many of these methods are, however, limited by the difficulty of computing the likelihood function, thus restricting their use to simple evolutionary scenarios and molecular

models. Additionally, even with ever-increasing computational power, these techniques cannot keep up with the demands of the large amounts of data generated by recently developed, high-throughput DNA sequencing technologies. Both of these factors have stimulated the development of new methods that approximate the likelihood (Csilléry *et al.*, 2010; Marjoram and Tavaré, 2006).

For these reasons, the parameter estimation problem is still a bottleneck and a challenging task of computational analysis of systems biology (Sisson *et al.*, 2007). Until now, none of the parameter estimation methods is effective in all cases and can overwhelm all the other methods. Instead, various methods have their advantages and disadvantages. Consequently, it is worthy to develop acceptably “good enough” methods within a given tolerance and time frame. In our work we consider one of the most recently developed technique - “Approximate Bayesian computation (ABC)” to calibrate parameters, which will be reviewed as follows.

1.4.2 Approximate Bayesian methods

For the case of parameter estimation when likelihoods are analytically or computationally intractable, approximate Bayesian computation (ABC) methods have been applied successfully (Beaumont *et al.*, 2002; Marjoram *et al.*, 2003) to bypass exact likelihood calculations by using summary statistics and simulations. ABC algorithms provide stable parameter estimates and are also relatively computationally efficient, therefore, they have been treated as substantial techniques for solving inference problems of various types of models that were intractable only a few years ago (Sisson *et al.*, 2007).

Generally, ABC algorithms can be classified into three broad categories. The first class relies on the basic rejection algorithm (Tavaré *et al.*, 1997; Pritchard *et al.*, 1999). Technical improvements of this basic scheme correct for the discrepancy

between the simulated and the observed statistics by using local linear or non-linear regression techniques (Beaumont *et al.*, 2002; Blum and François, 2010; Leuenberger and Wegmann, 2010). A second class, ABC-Markov chain Monte Carlo (MCMC) algorithms, explore the parameter space iteratively using the distance between the simulated and the observed summary statistics to update the current parameter values (Marjoram *et al.*, 2003; Wegmann *et al.*, 2009). The last one is inspired by Sequential Monte Carlo methods (SMC) (Liu, 2008), which approximates the posterior distribution using a large set of randomly chosen parameter values called “particles”. These particles are propagated over time by simple sampling mechanisms or rejected if they generate data that match the observation poorly. Ongoing work is seeking to improve the parameter space exploration and develop efficient sampling strategies that drive particles toward regions of high posterior probability mass (Sisson *et al.*, 2007; Toni *et al.*, 2009; Beaumont *et al.*, 2009).

For each of the above method, we start with a set of experimental data x and let θ be the parameter vector to be estimated. Assuming with a initial guess called prior distribution $\pi(\theta)$ and we want to approximate the posterior distribution $\pi(\theta|x)$. Details for algorithms are described as follows.

Generic form of ABC

All ABC algorithms obey the following major steps (Pritchard *et al.*, 1999).

1. Sampling step: sample a candidate parameter θ^* from the proposed prior distribution $\pi(\theta)$.
2. Simulation step: simulate the results x^* with the proposed model based on a conditional probability distribution $f(x|\theta^*)$.

3. Comparison step: compare the simulated data set x^* with the experimental data set x_0 and find the distance $d(x_0, x^*)$ between the two data sets in different ways.
4. Decision making step: choose a suitable tolerance or threshold value ϵ , then accept the sampled parameter θ^* if $d(x_0, x^*) \leq \epsilon$, otherwise, reject it and return to the first sampling step.

With sufficient amount of iterations for the above algorithm, we can obtain a set of estimated parameters from distribution $\pi(\theta|d(x_0, x^*) \leq \epsilon)$, which is an approximation for the posterior distribution $\pi(\theta|x_0)$. The difficulties here are to define a suitable distance function for calculating the difference and to choose an optimal tolerance value. If tolerance ϵ is sufficiently small, our obtained distribution will be a good approximation. However, it takes a long time to achieve a good approximation since it requires many samples before there are sufficient accepted samples to calculate the approximation. If tolerance is too large, we will obtain a distribution that maybe not satisfying for approximation.

ABC Monte-Carlo Markov Chain algorithm

Based on the generic form of ABC algorithm, many methods have been developed including ABC rejection sampler, which is a similar derivation as the above algorithm and ABC Monte-Carlo Markov Chain (ABC MCMC) (Marjoram *et al.*, 2003), which provides the following full steps.

1. Initialize $\theta_i, i = 0$
2. Sampling step: sample a candidate parameter θ^* from the proposed distribution $q(\theta|\theta_i)$.
3. Simulation step: simulate the results x^* with the proposed model based on a conditional probability distribution $f(x|\theta^*)$.

4. Comparison step: compare the simulated data set x^* with the experimental data set x_0 and find the distance $d(x_0, x^*)$ between the two data sets in different ways.
5. Decision making step: choose a suitable tolerance or threshold value ϵ , then if $d(x_0, x^*) \leq \epsilon$, set $\theta_{i+1} = \theta^*$ with probability

$$\alpha = \min\left[1, \frac{\pi(\theta^*)q(\theta_i|\theta^*)}{\pi(\theta_i)q(\theta^*|\theta_i)}\right]$$

and $\theta_{i+1} = \theta_i$ with probability $1 - \alpha$, otherwise, set $\theta_{i+1} = \theta_i$ and return to the first sampling step.

ABC MCMC algorithm solves the problem of long computing time with a badly chosen prior distribution that is far away from posterior distribution using ABC rejection sampler algorithm. However, as ABC MCMC introduced a concept of acceptance probability during the decision making step, then candidate parameters must meet two criteria. This may result in getting stuck in the regions of low probability for the chain and we may never be able to get a good approximation.

ABC Sequential Monte-Carlo algorithm

Instead of having one parameter vector at a time, we sample from a pool simultaneously treating each parameter vector as a particle. The algorithm starts with sampling a pool of N particles for parameter vector θ through prior distribution $\pi(\theta)$. The sampled particle candidates $\theta_1^*, \dots, \theta_N^*$ will be chosen randomly from the pool and we will assign each particle a corresponding weight w to be considered as the sampling probability. For the first iteration, we assume for each sampled particle, it has an equally weight of $\frac{1}{N}$. A perturbation and filtering process following through a transition kernel $K(\cdot|\theta^*)$ finds the particles θ^{**} . Similarly with θ^{**} , data x^* can be simulated and compared with experimental data x .

Many algorithms have been developed using the particle filtering, such as partial rejection control, population Monte-Carlo and sequential Monte-Carlo (SMC) (Del Moral *et al.*, 2006; Sisson *et al.*, 2007). Each of them differs in how the formation weight w that are assigned and the transition kernels $K(\cdot|\theta^*)$ they choose. We will only present the basic ABC SMC algorithm (Toni *et al.*, 2009), which is a special case of sequential importance sampling (SIS) algorithm here (Del Moral *et al.*, 2006).

1. Initialize $\epsilon_1, \dots, \epsilon_T$, start with iteration $t = 1$ as well as the particle indicator $i = 1$.
2. Sampling step: If $t = 1$, sample θ^{**} from the a proposed prior distribution $\pi(\theta)$. Else, sample from the previous population $\theta_{t-1}^{(i)}$ with weight w_{t-1} and perturb the particle to obtain $\theta^{**} \sim K_t(\theta|\theta^*)$.
3. Simulation step: simulate the results x^* with the proposed model based on a conditional probability distribution $f(x|\theta^{**})$.
4. Comparison step: compare the simulated data set x^* with the experimental data set x_0 and find the distance $d(x_0, x^*)$.
5. Decision making step: if $d(x_0, x^*) \leq \epsilon$, set $\theta_t^{(i)} = \theta^{**}$, find the weight for particle $\theta_t^{(i)}$,

$$w_t^{(i)} = \begin{cases} 1 & \text{if } t = 1, \\ \frac{\pi(\theta_t^{(i)})}{\sum_{j=1}^N w_{t-1}^{(j)} K_t(\theta_t^{(j)}|\theta_t^{(i)})} & \text{if } t > 1; \end{cases}$$

set $i = i + 1$ and return to the first sampling step.

ABC SMC is a promising tool for reliable parameter inference and model selection for models of dynamical systems that can be efficiently simulated. Owing to its

simplicity and generality, ABC SMC, unlike most other approaches, can be applied without any change in both deterministic and stochastic contexts (including models with time delay) (Toni *et al.*, 2009). Moreover, in the context of hypothesis testing, the Bayesian perspective (Cox and Hinkley, 1979; Robert and Casella, 2013) has a more intuitive meaning than the corresponding frequentist point of view.

1.4.3 Parameter identification

Parameter identification is an important part of network inference and the prediction of network behaviour in time (Omony, 2014). While parameter estimation is not feasible, we can still study the system by assessing which parameters significantly affect the outputs or measured variables of interest following a network perturbation with some stimuli (Liang *et al.*, 1998). Conventionally parameter sensitivity analysis is used as a tool for analysis and design in engineering systems theory. Although it has mostly been applied extensively in physical systems rather than biological systems, its use in the latter has increased recently, especially in the study of complex networks. By using parameter sensitivity analysis, once the most influential parameters are identified, the correlation matrix between the parameters is then investigated. Thereafter, the least sensitive parameters can be left out of a model thereby reducing the model complexity but still retaining its explanatory power (Erban *et al.*, 2006). This technique is especially essential for large networks with thousands of genes, the number of differential equations required to describe a particular system becomes huge (Bornholdt, 2005), which implies an increased number of kinetic parameters, e.g. mRNA production and decay rates and Hill constants. In principle, using parameter sensitivity analysis, some of these parameters can be coalesced or dropped from the model, leaving a simpler yet still powerful model to describe the network dynamics.

1.5 Objectives

Building a systematic understanding of biological networks currently has become one of the hot topics for both biologist and mathematicians facing modern biology. An increasing number of researches in recent years showed that computational modelling has provided deep understanding as well as experimentally testable predictions regarding cellular dynamics at a system level. However, experimental discoveries regarding the discrete processes inside the cell have increasingly posed great challenges to mathematical modelling and computer simulations. In order to understand the functioning of organisms on the molecular level, we need to understand how the genes express themselves, when and in which organisms. The regulation of gene expression is achieved through networks of interactions between DNA, RNA, proteins and small molecules. In this thesis, I am focused on mathematical modelling for particular biological systems including construction of structures and parameter estimations during modelling process. The following paragraphs will describe how we came to these two points from our initial objectives, the corresponding methods that have been developed and how the objectives are achieved step by step.

While modelling various complex biological systems, the development of simple mathematical models for representing complicated real-life chemical reaction systems has been a fundamental issue in computational biology and bioinformatics. In particular, the accurate description of chemical events of multi-step chemical reactions has been regarded as an essential problem in chemistry and biophysics. In recent years, a number of modelling approaches have been attempted to use simplified models to describe multi-step chemical reactions accurately. However, more sophisticated modelling methods are strongly required in order to provide more accurate description of complex biochemical systems in an efficient way. In addition to building mathematical models for such biological systems, one of

the major challenges in systems biology is how to infer unknown parameters in mathematical models based on the experimental data sets, in particular, when the data are sparse and the regulatory network is stochastic. Several inference methods including optimization methods and Bayesian inferences have been approached for estimating unknown parameters during the last decade. Among the vast range of inference techniques, Bayesian inference methods have more power to solve problems when modelling biological systems where molecular species are present in low copy numbers and noise exists (Raj and van Oudenaarden, 2008; Wilkinson, 2007). Another challenging issue in mathematical modelling is to study the influence of parameter variations on the system property. Robustness and sensitivity properties are two major measurements to describe the dynamic property of the system against the variation of model parameters. For stochastic models of discrete chemical reaction systems, although these two properties have been studied extensively, no work has been done so far to investigate these two properties together.

In detail, our objectives are listed as follows.

1. Development of novel mathematical models for multi-step chemical reaction systems, which includes
 - Study of dynamics for one of the most typical multi-step chemical reaction systems - “mRNA degradation process”;
 - Development of two-variable mathematical model which simplifies the traditional fully described ordinary differential equation model;
 - Development of a simplified mathematical model with state-dependent time delay.
2. Application and development of approximate Bayesian computation (ABC):

- Apply general ABC methods to estimate parameters for stochastic biological systems;
 - Develop a novel algorithm using simulated likelihood density in the framework of approximate Bayesian computation for parameter estimations.
3. Application and development for parameter identification methods:
- Develop a new framework for the method to analysis sensitivity;
 - Combine sensitivity analysis with study of robustness property for a biological system to understand the influence of system input -“parameter variance”.
4. Network inference for a P53 gene network:
- Propose an ntegrated method that combines a top-down approach and a bottom-up approach to investigate the dynamics of regulatory networks through high throughput experimental data, such as microarray gene expression profiles;
 - Apply model simulation error, Akaike’s information criterion, parameter identifiability and robustness property as criteria to select the optimal network.

1.6 Thesis Outline

This thesis addresses the above four objectives outlined above in four individual publications and three other unpublished papers, which are included between the thesis introduction and conclusion chapters. The contents in each chapter and their connections are described as below.

Chapter 1 of the thesis is an introductory chapter highlighting the motivations, difficulties of mathematical modelling and parameter inference methods, discussion of literature and the objectives, which are briefly introduced above.

Chapters 2 and 3 of the thesis were our first trial to tackle first part of objective 1. To describe stochastic process in regulatory networks, the stochastic simulation algorithm (SSA) has been widely used to simulate chemical reaction systems. This method represents an essentially exact procedure for modelling reaction systems in which the molecular population of some critical reactants is relatively small. This modelling framework is based on the assumption that all biochemical reactions are instantaneous events. However, biological systems are complex; thus this assumption is not adequate for describing the complex dynamics by using the simplified mathematical models. Multi-step chemical reactions were traditionally simplified into a one-step reaction and this usually cannot provide concrete description of the dynamics of multi- step reactions. In Chapter 2 (Wu *et al.*, 2012) and Chapter 3 (Wu *et al.*, 2013b), we successfully built a two-variable model, which introduces a new concept regarding the location of molecules in the multi-step reactions in order to simplify chemical events of multi-step reaction. The efficiency of the proposed two-variable model is demonstrated by the realization of mRNA degradation process based on the experimentally measured data, which is shown in these two chapters.

Chapter 4 (Wu and Tian, 2015) continues addressing the first objective. In the current modelling approaches with time delays, it is widely assumed that time delay is either a constant or a random variable that follows a given distribution. To model chemical reaction systems in a manageable way to further fulfil our objective 1, we consider that time delay is dependent on the system state, rather than to be a constant. This consideration is reasonable because the waiting times of all chemical reactions are system state dependent. In fact, the state-dependent

time delay has been studied in a number of research areas such as optimal control, which is discussed in Chapter 4.

Approximate Bayesian computation (ABC) algorithms provide stable parameter estimates and are also relatively computationally efficient, therefore, they have been treated as substantial techniques for solving inference problems of various types of models that were intractable only a few years ago (Sisson *et al.*, 2007). We extended our research based on the generic form of ABC algorithms, studying several valid ABC techniques such as ABC Markov chain Monte Carlo (MCMC) etc. (Golightly and Wilkinson, 2011) and proposing our own ABC algorithms using simulated transitional density function as the objective function and different strategies for defining errors. Detailed description of the proposed algorithms can be found in the Chapter 5 and 6, which address our objective 2.

The third objective is presented with Chapter 7 (Wu *et al.*, 2015) , which focused on a proposed new framework to study the sensitivity and robustness properties for a biological system simultaneously. Using stochastic model as the test system, we aim at identifying key coefficients that have the most influence on the dynamics of the network. Numerical results suggest that the proposed framework is an efficient approach to study the sensitivity and robustness properties of biological network models.

Chapter 8 (Wang *et al.*, 2015) is to address our final objective, which extends to network inference for the dynamics of protein-gene interactions with human gene P53 case. Along with the parameter inference problems, a new integrated method is proposed by combining a top-down approach using probability graphic models and a bottom-up approach using differential equation models are used to predict the network structure of DNA repair pathway that is regulated by the p53 protein and study the detailed genetic regulations. Overall, the new integrated method is

a promising approach for investigating the dynamics of genetic regulation as well as for parameter estimation.

Chapter 9 presents some concluding remarks with an overview of the results, the contributions of the thesis and directions for future work.

Since this is a "Thesis by Publication" which consists of a new introduction and conclusion with published papers in between, unfortunately, it has inevitably created some amount of repetition among chapters, especially chapters (2,3) and (5,6) that are pairs of conference proceeding papers and journal papers. For the sake of thesis unity, all the references of publications are located in a single Bibliography after chapter 9 and acknowledgements in the publications are covered by the thesis Acknowledgement.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 2

Declaration by candidate

In the case of Chapter 2, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Developed, established and verified the method Wrote programming codes and the article	90%

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Kate Smith-Miles	Provided helpful guidance and proofreading	
Tianhai Tian	Provided helpful guidance and editorial work	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature		Date 29/04/15
Main Supervisor's Signature		Date 29/04/15

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 2

A Two-variable Model for Stochastic Modelling of Chemical Events with Multi-step Reactions

Chapter 2 is based on the article Wu Q, Smith-Miles K, Tian T. 2012. A two-variable model for stochastic modelling of chemical events with multi-step reactions. In: Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on. pp. 1–6, doi: 10.1109/BIBM.2012.6392681.

Abstract. *The development of simple mathematical model for representing complicated real-life chemical reaction systems has been a fundamental issue in computational biology and bioinformatics. In particular, the accurate description of chemical events with multi-step chemical reactions has been regarded as an essential problem in chemistry and biophysics. To model chemical reaction systems in a manageable way, multi-step chemical reactions were normally simplified into a one-step reaction. In recent years, a number of modelling approaches have been attempted to use simplified model to describe multi-step chemical reactions accurately. In this work, we proposed a two-variable model to describe chemical events with multi-step chemical reactions. We introduced a new concept to represent the location of molecules in the multi-step reactions, and use it as the second indicator of the system dynamics. The accuracy of the proposed new model was evaluated via using a deterministic model. The proposed model has been applied to study the mRNA degradation process. Numerical simulations of the designed simplified models matched the simulations of multi-step chemical reactions very well.*

Keywords. *Stochastic modelling, multi-step reactions, mRNA degradation.*

References are considered at the end of the thesis.

Chapter 2

A Two-variable Model for Stochastic Modelling of Chemical Events with Multi-step Reactions

2.1 Introduction

Recent advances in computational biology and bioinformatics have provided a variety of mathematical models to describe complex chemical reaction systems inside the cell. There has been amount of evidence showing that mathematical modelling is a powerful and predictive tool for exploring the dynamic properties of genetic regulatory networks, cell signalling transduction pathways and metabolic pathways (Lewis, 2008; Tomlin and Axelrod, 2007). In spite of the substantial progress, there are still a number of fundamental issues that need to be addressed imperatively. Among them, the accurate description of chemical events with multi-step chemical reactions has been regarded as a central problem in chemistry and biophysics (Zhou and Zhuang, 2007). There are many biochemical events involving multi-step chemical reactions. One of the most well-known example

is gene expression which usually involves a large number of steps including transcription, RNA processing, DNA translation and messenger RNA (mRNA) degradation, which can all be considered as multi-step reaction systems. In particular, transcription is a multi-step process consisting of initiation, elongation and termination phases; and elongation is a sequence of reactions that occur at each elongation step for RNA polymerase passing through the DNA. The multi-step reaction processes also exist in other areas such as organic chemistry and biophysical chemistry (Branz, 1996). For example, an ion channel may change its conformation through multi-step allosteric transitions (Qin and Li, 2004). Therefore to accurately describe chemical events with multi-step reactions is a critical step in the development of mathematical models for characterizing complex biological systems.

To model chemical reaction systems in a manageable way, multi-step chemical reactions were traditionally simplified into a one-step reaction. For example, it was a widely used approach to use first order reactions to describe the degradation process of mRNA or protein. Since the simplified one-step reaction cannot provide concrete description of the dynamics of multi-step reactions, recently chemical reactions with time delay have been used to describe the multi-step chemical events or slow reactions more accurately (Monk, 2003). To address the coupling of intrinsic noise in biochemical reactions with delays, a new method called Delay Stochastic Simulation Algorithm (DSSA) was proposed by introducing time delay into the Stochastic Simulation Algorithm (SSA) (Barrio *et al.*, 2006). Unlike the classic SSA, which assumes instantaneously biochemical reaction systems in the model, the DSSA was designed to characterize chemical systems with both fast and slow reactions. In fact the so-called slow reaction in most cases is a simplified version of the multi-step reactions. This delayed method has been applied for many physical and biological systems. For example, Barrio *et al.* (2006) applied the DSSA to mimic delays associated with transcription and translation

and successfully explained the process for the regulation of Hes1 gene. These simulation methods have also been used to successfully validate stochastic models of biological systems with slow reactions. Recently the work done by Mier-y Terán-Romero *et al.* (2010) opened some new aspects for application of time delays in biological systems. They presented the developed time delay models for protein translation based on the partial differential equation (PDE) models and obtained a good agreement between the time delay model and mechanistic models, which allows us for further study of formulation of time delay models of coupled template polymerization process in modelling of genetic networks (Mier-y Terán-Romero *et al.*, 2010). Other modelling techniques proposed recently include the slow-scale linear noise approximation and the stochastic quasi-steady-state assumption (Thomas *et al.*, 2012; Srivastava *et al.*, 2011).

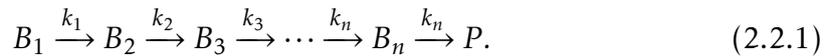
The degradation process of mRNA illustrates a typical system with multi-step reactions and is also an important step in the regulation of gene expression (Mitchell and Tollervey, 2001). Over the past decade, mRNA degradation has been studied deeply, but there still exist problems unsolved with respect to the enzymes, pathways and regulation of mRNA degradation including the role of P-bodies etc. (Garneau *et al.*, 2007; Shyu *et al.*, 2008; van Hoof and Parker, 2002). Based on the detailed process including poly(A) tail shortening, decapping and digestion, mathematical models have been designed for understanding the dynamics of mRNA degradation, including a linear multi-component model that was designed to investigate the mRNA degradation problem as well as the nonsense-mediated decay of mRNA molecules in yeast (Cao and Parker, 2001, 2003). This model includes 23 first-order reactions that describes transcription, decapping, poly(A) shortening, translocation and as well as digestion process. It is the first detailed deterministic model that studies mRNA degradation. Simulation results suggested that the widely used mRNA half-life, obtained by using the first order reaction, underestimated the averaged life-span of mRNA molecules and also half-life

is an important factor for determining the different steps in the degradation pathway. With robustness analysis, it showed that the change of deadenylation rate might lead to great variations in mRNA copy numbers. To interpret the complex reactions in this detailed mathematical model, a multi-step reaction model was proposed recently by using a chain of 11 chemical reactions (Tian, 2014). Numerical simulations suggested that this simplified model gave a very good approximation to the original detailed model with 23 chemical reactions.

To further simplify the degradation process of mRNA, another approach used time delay to represent the total time required in the multi-step reactions (Tian, 2014). The simplified stochastic model with time delay was also adopted for studying the degradation process of mRNA molecules. Numerical results showed that the simple first-order reaction models could not approximate the detailed degradation process precisely (Tian, 2014). And even with delay introduced, it still remains a challenge to represent the chemical events with multiple small step reactions accurately. Therefore, instead of using time delay to represent the missing intermediate reactions in the one-step reaction, we here introduce another modelling method by introducing a new concept, which is termed as the length of a molecule to represent the location of that molecule in the multi-step reactions, and use it as the second indicator of the system dynamics. The following sections are organized as follows. Section 2.2 will introduce the new modelling approach with two variables for describing chemical events with multi-step reactions. The accuracy of the proposed new model is evaluated with a deterministic model in section 2.3. Section 2.4 of this paper studies the mRNA degradation process using our new modelling approach.

2.2 New Modelling Method for Multi-step Reaction System

Let us consider the following chemical events with multi-step chemical reactions, which is adapted from the theoretical model studied by Zhou and Zhuang (2007):



In this system, any molecule that starts from the “ B_1 ” state has to experience $(n-1)$ intermediate states B_2, \dots, B_n before it is turned to a “ P ” state. The molecule P may be the product of this multi-step process. It may also represent the degradation process when $P = ()$.

We denote X as the total copy number of molecules B_i

$$X = \sum_{i=1}^n [B_i].$$

In addition, according to the distance to the final product, for each B_i molecule, we define a corresponding length of $n - i + 1$, therefore the total length of all molecules is given by

$$L = \sum_{i=1}^n (n - i + 1)[B_i].$$

The proposed new model was considered in the following way. When a reaction occurs, the total length will decrease by one while the total copy number of molecules may remain the same if the reaction is one of the first $(n-1)$ steps or decrease by one if the reaction is the last step. Therefore, the two-variable reaction model can be structured via two types of reactions:



representing reactions for $B_i \xrightarrow{k_i} B_{i+1}$, and



representing the reaction for $B_n \xrightarrow{k_n} P$.

After suggesting two-variable reaction model, the SSA method that is the basic approach for various forms of modelling chemical systems will be applied to simulate the new model. It is described by the following algorithm.

Algorithm I

1. Based on the total molecule number X and total length L , we can calculate the propensity function

$$a_0 = kX,$$

where k is the harmonic mean of the rate constants

$$k = \frac{n}{\frac{1}{k_1} + \dots + \frac{1}{k_n}}. \quad (2.2.4)$$

2. Determine the step size for the next reaction

$$\tau = \frac{1}{a_0} \ln \frac{1}{r_1},$$

where $r_1 \sim U(0, 1)$.

3. Generate a sample $r_2 \sim U(0, 1)$ to determine which reaction from reactions (2.2.2) and (2.2.3) will occur,

$$(X, L) = \begin{cases} (X, L - 1) & \text{if } r_2 > f, \\ (X - 1, L - 1) & \text{if } r_2 < f, \end{cases}$$

where f is the probability of the firing of the last reaction and then the system is updated.

4. Go back to step 1.

The key question remaining in the proposed model is to define a proper probability function f to describe the firing of the last reaction. It is clear that this probability should depend on the values of total molecule number X , total length L and the number of reactions n . Our initial attempt suggested that it could be difficult to find an analytical expression of the probability function $f(X, L, n)$.

In this work, we proposed to use the following expression, given by

$$\text{Type I:} \quad f(X, L, n, q) = 1 - \left(\frac{L - X}{X(n - 1)} \right)^q, \quad (2.2.5)$$

and an alternative expression is

$$\text{Type II:} \quad f(X, L, n, q) = \left(1 - \frac{L - X}{X(n - 1)} \right)^q. \quad (2.2.6)$$

The aim for this work is to test the feasibility of the two proposed functions f , find for the optimal q value under various simulation methods and apply the new modelling method to biological systems such as the process of mRNA degradation, which will be introduced in the following sections 2.3 and 2.4.

2.3 Ordinary Differential Equation Model

After we find the linear relation between the optimal value of q and the number of reactions n through the probability simulations, we next studied the corresponding ordinary differential equation (ODE) model to test the feasibility of the approximation probability function. Solving a set of ODEs numerically is another

common way for describing the system with a set of chemical reactions. From multi-step chemical reaction system (2.2.1), the ODE model is formed as follows:

$$\begin{aligned}\frac{dB_1}{dt} &= -k_1B_1, \\ \frac{dB_2}{dt} &= k_1B_1 - k_2B_2, \\ &\vdots \\ \frac{dB_n}{dt} &= k_{n-1}B_{n-1} - k_nB_n, \\ \frac{dP}{dt} &= k_nB_n.\end{aligned}\tag{2.3.1}$$

We can calculate the total molecule number X as

$$X = B_1 + \cdots + B_n,$$

and the total length of the molecules L is

$$L = B_n + 2B_{n-1} + \cdots + nB_1.$$

By adding up all the ODEs, then we have a new set of ODEs for the total molecule number X and total length L , given by

$$\begin{aligned}\frac{dX}{dt} &= -kB_n, \\ \frac{dL}{dt} &= -kX,\end{aligned}\tag{2.3.2}$$

where k is the harmonic mean of the rate constants (2.2.4). Note that in the (2.3.2), kB_n can be approximated with the probability function f as they all act as the probability for the occurrence of last step reaction. Thus in this work we proposed the following ODE model to represent the chemical events with

multi-step chemical reactions

$$\begin{aligned}\frac{dL}{dt} &= -kX, \\ \frac{dX}{dt} &= -kX\left(1 - \frac{L-X}{X(n-1)}\right)^q.\end{aligned}\tag{2.3.3}$$

From the original ODEs (2.3.1), we can find the exact solutions for B_i with the given initial conditions $B_i(0)$ and rate constants k_i under a certain long enough time frame. The approximated ODEs (2.3.3) will be solved numerically for various q values under the same conditions.

Simulations were operated with conditions of $n = [5\ 10\ 15]$ and $X = [5\ 10\ 50\ 100\ 200\ 500]$ respectively, i.e. for each n value we use the function `ode23s` in MATLAB to solve the system (2.3.3) for different cases with various initial X values. With the simulation results, Fig. 2.1 (A) was plotted, which show some patterns such that optimal q increases when number of chemical reactions n increases and it does not fluctuate significantly with various initial X values.

A linear regression for the averaged optimal q versus n can be revealed from Fig.2.1 (B), the equation is described in the following form

$$\bar{q} = 0.4570n + 0.8567.\tag{2.3.4}$$

One of the important findings is that the value of q is dependent on the number of reactions n , but independent on the total copy number X . In addition, when q is close to the optimal q value, the difference between the error of optimal approximation and that using q is quite small. As a general rule, the value of q can be approximated with $n/2$.

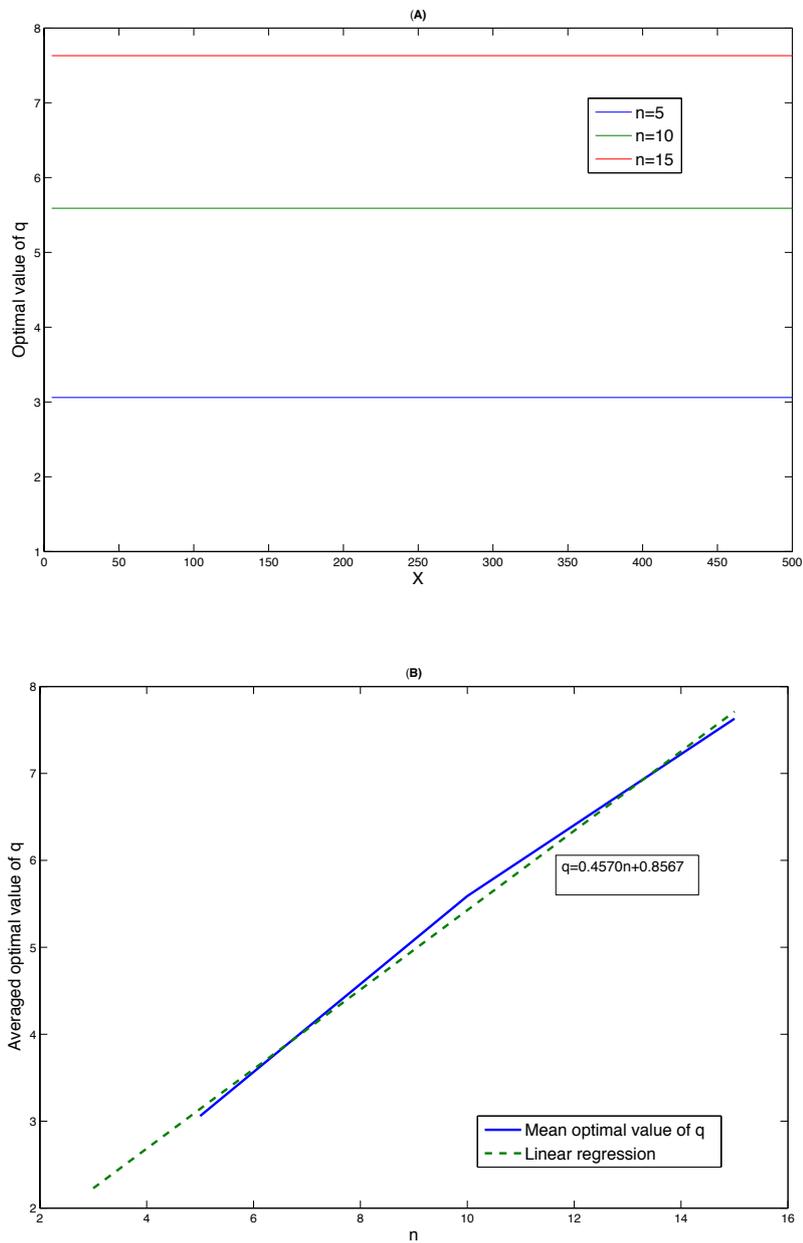


Figure 2.1: Simulation results from the ODE model: (A) for the optimal q value against different X , (B) averaged optimal values for q against n .

With these findings of linear relationship between q and n , we can rewrite the probability function f in terms of L , X , and n such as

$$f(L, X, n) = \left(1 - \frac{L - X}{X(n - 1)}\right)^{0.4570n + 0.8567}. \quad (2.3.5)$$

2.4 Application to mRNA Degradation

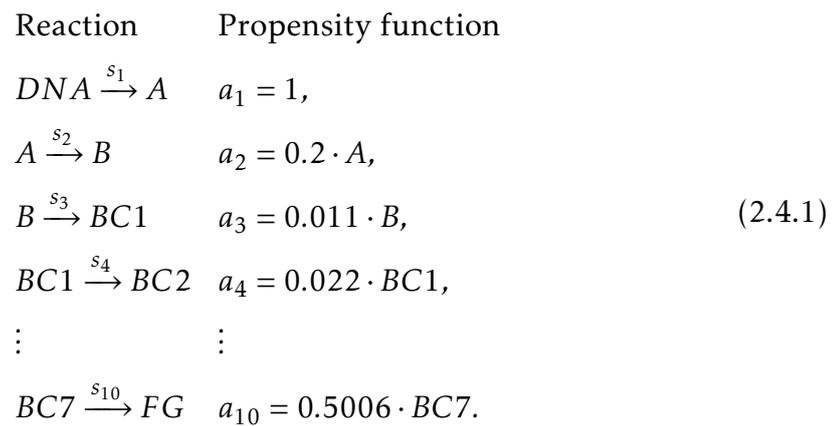
Based on the linear multi-component model studied by Cao and Parker (2001), a simplified mathematical model to represent the mRNA decay was proposed, which is presented by the following Table 2.1 (Tian, 2014). It is assumed that the gene transcription is a zeroth-order process, which is given by reaction S_1 under a rate constant of 1. And then the mRNA molecules species A in the nucleus will translocate into the cytosol as species B via reaction S_2 under a rate constant of 0.2. Different from the original model, it was suggested that species B would start the poly(A) shortening process through reactions S_3, \dots, S_9 with various rates instead of undergoing decapping reaction, 5'-to-3' / 3'-to-5' exonucleolytic degradation or digestion processes. Reaction S_{10} is a further exonucleolytic degradation to trim the mRNA with a poly(A) tail length of zero to produce species FG , which will be degraded in the end by reaction S_{11} . Since the fragment product (FG) is not a functional mRNA, we excluded reaction S_{11} for our consideration.

To apply our proposed two-variable model to this mRNA degradation process, the following reaction model is built by constructing realization of $\mathbf{X} = (A, B, BC1, \dots, BC7)$. Each reaction has its corresponding propensity function

Table 2.1: Reactions and kinetic rates of the simplified stochastic model. The rate constants s_i are in the unit of 1/sec.

	Reaction	Rate constant s_i	Comment
S_1	$DNA \rightarrow A$	1	transcription
S_2	$A \rightarrow B$	0.2	transport
S_3	$B \rightarrow BC1$	0.011	full-length 70A-60A
S_4	$BC1 \rightarrow BC2$	0.022	full-length 60A-50A
S_5	$BC2 \rightarrow BC3$	0.022	full-length 50A-40A
S_6	$BC3 \rightarrow BC4$	0.022	full-length 40A-30A
S_7	$BC4 \rightarrow BC5$	0.022	full-length 30A-20A
S_8	$BC5 \rightarrow BC6$	0.023	full-length 20A-10A
S_9	$BC6 \rightarrow BC7$	0.0099	full-length 10A-0A
S_{10}	$BC7 \rightarrow FG$	0.5006	fragment production
S_{11}	$FG \rightarrow ()$	0.00066	fragment degradation

shown as below:



The total mRNA molecule number X and total length L can be calculated as

$$X = A + B + BC1 + \dots + BC7,$$

$$L = 9A + 8B + 7BC1 + \dots + BC7.$$

The SSA that generates a trajectory of the system step by step instead of following the time evolution of the probabilities is used here for the simulation. In each step, the SSA starts from its current system state $\mathbf{x}(t) = \mathbf{x}$ and examine itself two questions: When will the next reaction occur and which reaction will it be? Gillespie derived the formula for answering these two questions by studying the joint probability density function $p(\tau, j|\mathbf{x}; t)$, where τ is the time interval for next reaction to occur. And for each reaction R_j , the propensity function $a_j(\mathbf{x})$ is defined by a given state $\mathbf{x}(t) = \mathbf{x}$ and the value of $a_j(\mathbf{x})dt$ that represents the probability of one reaction will occur somewhere during the infinitesimal time interval $[t, t + dt)$ (Gillespie, 1977). The SSA is an exact procedure for generating the time and index of the next occurring reaction according its current state and the propensity functions, which are defined as

$$a_0(\mathbf{x}) = \sum_{j=1}^M a_j(\mathbf{x}).$$

Also the time interval τ can be obtained with

$$\tau = \frac{1}{a_0(\mathbf{x})} \ln \frac{1}{r_1},$$

where $r_1 \sim U(0, 1)$.

With the SSA simulations for above model (2.4.1), we will have the exact solutions for this mRNA degradation problem. On the other hand, using the proposed two variable chemical reaction systems and the approximated function (2.3.5) we finalized, we can set up a simpler model for the same mRNA degradation problem through constructing realization $\mathbf{X} = (L, X)$. With $n = 9$, each chemical reactions

with its corresponding propensity function are described as

Reaction	Propensity function
$DNA \xrightarrow{k_1} (9L, X)$	$a_1 = k_1,$
$(X, L) \xrightarrow{k} (X, L - 1)$	$a_2 = k \cdot X \cdot (1 - f(L, X, 9)),$
$(X, L) \xrightarrow{k} (X - 1, L - 1)$	$a_3 = k \cdot X \cdot f(L, X, 9).$

Note that here $k_1 = s_1 = 1$ as it's the rate for producing mRNA species. The rate constant k can be calculated from (2.2.4) with (S_2, \dots, S_{10}) , given as

$$k = \frac{9}{\sum_{i=2}^{10} \frac{1}{s_i}} = 0.0212 \quad (2.4.2)$$

Initial conditions we took here are $\mathbf{X} = [10 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$ for the exact SSA and $\mathbf{X} = [90 \ 10]$ for the approximated SSA. We carried out two numerical tests.

When $s_1 = k_1 = 1$, we notice that the rate of last-step degradation is $s_{10} = 0.5006$ and the synthesis rate for mRNA species A is $s_1 = k_1 = 1$, which is greater than s_{10} . Fig. 2.2 shows that both X and L will become steady in the long run as the equilibrium achieves. Fig. 2.2 (A) and (C) represent an example of three simulations of X and L over a time period of 1,300 seconds from the exact results derived from the detailed multi-step reaction model, while Fig. 2.2 (B) and (D) show the three approximated simulation results for X and L over the same amount of time. By taking the average over the 10,000 simulations, the averaged values for X and L can be compared for both models shown by Fig. 2.2 (E) and (F) respectively. They reveal that the approximated solutions approach the exact simulations very well.

Instead of having non-zero rate of reaction S_1 , we also simulated the two models with the zero rate of mRNA synthesis, which means that there will be having

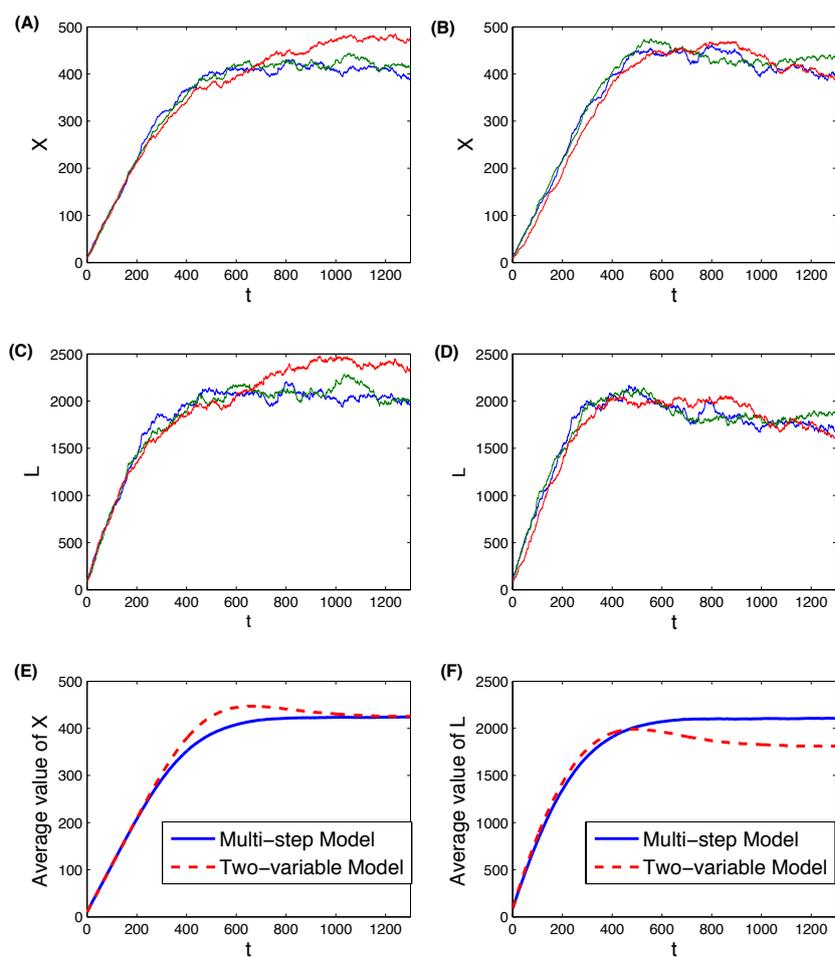


Figure 2.2: The SSA simulation results with $s_1 = k_1 = 1$: (A,C) three simulations of X and L values over t for the detailed multi-step reaction model, (B,D) three simulations of X and L values over t for approximated model with two variables, (E) the mean value of X with 10,000 simulations for both models, (F) the mean value of L with 10,000 simulations for both models.

no more further production of new mRNA species A molecules adding into the reaction system. Therefore, unlike the previous case, L and X will both decrease and tend to 0 eventually, which can be revealed from Fig. 2.3. And with this numerical test, it also shows that the approximated solutions are close to the exact results. Hence with this application, we found that the approximated SSA using the proposed two-variable model indeed creates a good approximation for the exact model. It further confirms that the form of q achieved before is a good approximation. In addition, we found that the computing time taken for generating simulation results using approximated two-variable model is much shorter than the ones using the detailed multi-step reaction method.

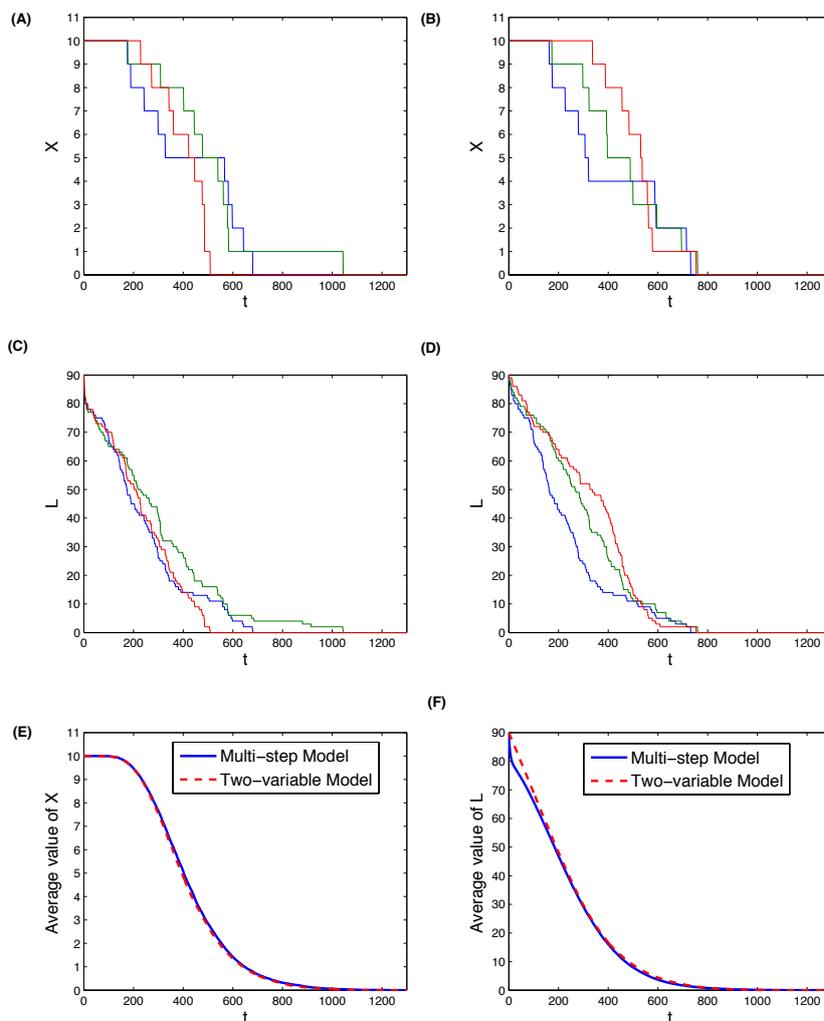


Figure 2.3: The SSA simulation results with $s_1 = k_1 = 0$: (A,C) three simulations of X and L values over t for the detailed multi-step reaction model, (B,D) three simulations of X and L values over t for approximated model with two variables, (E) the mean value of X with 10,000 simulations for both models, (F) the mean value of L with 10,000 simulations for both models.

2.5 Conclusion

In this work, we have proposed a new model to describe chemical events with multi-step chemical reactions. This represents a major step in designing simplified mathematical model to represent complex chemical reactions systems, which is a fundamental issue in computational biology and bioinformatics. In addition to the total molecule number, we proposed to use the length of a molecule to represent its location in the multi-step chemical reactions. We used the ODE model to find the optimal value in the non-linear function via comparison of the simulations derived from detailed multi-step chemical reaction model and our proposed two-variable model. Our designed model has been successfully applied for the stochastic simulations of the mRNA degradation process. Numerical simulations of the designed simplified models match the simulations of the stochastic model with multi-step chemical reactions very well.

However, there are still a number of challenging issues that require further research to address. The core of the proposed new model is a non-linear function that is designed to approximate the probability of the firing of the last chemical reaction. On top of that, more accurate information regarding the probability will clearly lead to more sophisticated stochastic models to describe chemical events with multi-step reactions. In addition, the derived relationship between the optimal value in the non-linear function and the key parameters of the multi-step reactions should be further validated by stochastic simulations that is a more appropriate approach to describe biological systems with small copy numbers of molecules. Finally we discussed the mRNA degradation process in this work by adding the synthesis of mRNA molecules into the multi-step reaction system. It is expected that the proposed two-variable model will be incorporated into more complex biological systems including genetic regulatory networks, telomere length regulation as well as cell differentiation and death.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 3

Declaration by candidate

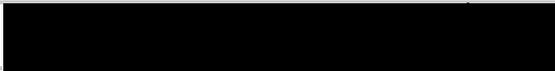
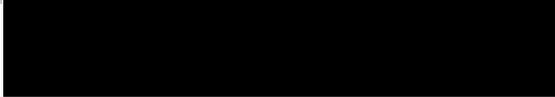
In the case of Chapter 3, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Developed, established and verified the method Wrote programming codes and the article	85%

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Kate Smith-Miles	Provided helpful guidance and proofreading	
Tianshou Zhou	Provided helpful guidance and proofreading	
Tianhai Tian	Provided helpful guidance and editorial work	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature		Date 29/04/15
Main Supervisor's Signature		Date 29/04/15

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 3

Stochastic Modelling of Biochemical Systems of Multi-step Reactions using a Simplified Two-Variable Model

Chapter 3 is based on the article Wu Q, Smith-Miles K, Zhou T, Tian T. 2013b. Stochastic modelling of biochemical systems of multi-step reactions using a simplified two-variable model. BMC Systems Biology 7(4): S14, doi: 10.1186/1752-0509-7-S4-S14, URL <http://dx.doi.org/10.1186/1752-0509-7-S4-S14>. Abstract.

Background: *A fundamental issue in systems biology is how to design simplified mathematical models for describing the dynamics of complex biochemical reactions systems. Among them, a key question is how to use simplified reactions to describe the chemical events of multi-step reactions that are ubiquitous in biochemistry and biophysics. To address this issue, a widely used approach in literature is to use one-step reaction to represent the multi-step chemical events. In recent years, a number of modelling methods have been designed to improve the accuracy of the one-step reaction, including reactions with time delay. However, our recent research results suggested that there are still deviations between the dynamics of delayed reactions and that of the multi-step reactions. Therefore, more sophisticated modelling methods are needed to accurately describe the complex biological systems in an efficient way.*

Results: *This work designs a two-variable model to simplify chemical events of multi-step reactions. In addition to the total molecule number of a species, we first introduce a new concept regarding the location of molecules in the multi-step reactions, which is the second variable to represent the system dynamics. Then we propose a simulation algorithm to compute the probability for the firing of the last step reaction in the multi-step events. This probability function is evaluated using a deterministic model of ordinary differential equations and a stochastic model in the framework of the stochastic simulation algorithm. The efficiency of the proposed two-variable model is demonstrated by the realization of mRNA degradation process based on the experimentally measured data.*

Conclusions: *Numerical results suggest that the proposed new two-variable model produces predictions that match the multi-step chemical reactions very well. The successful realization of the mRNA degradation dynamics indicates that the proposed method is a promising approach to reduce the complexity of biological systems.*

References are considered at the end of the thesis.

Chapter 3

Stochastic Modelling of Biochemical Systems of Multi-step Reactions using a Simplified Two-Variable Model

3.1 Introduction

The advances in systems biology have raised the importance of quantitative methods for studying various systems in molecular biology. In recent years, various research methods, including mathematical modelling, statistical analysis, computer simulation and visualization, have been employed to investigate the dynamic or statistical properties of regulatory networks. In particular, mathematical models have been widely used to describe the dynamics of complex systems inside the cell, including genetic regulatory networks, cell signalling transduction pathways and metabolic pathways (Lewis, 2008; Tomlin and Axelrod, 2007). However, these substantial progresses have further raised a number of fundamental and challenging issues that require to be addressed imperatively.

One of the major challenges in systems biology is how to use simple mathematical models to describe complex biological systems. To address this issue, a number of modelling techniques have been designed. Among them, a widely used approach is to use one-step reaction to represent multi-step reactions, which is also called slow reaction. This technique is very important because recent theoretical and experimental studies have shown that a wide variety of biochemical events involve multi-step reactions (Zhou and Zhuang, 2007). Perhaps the most important example of multi-step reactions is transcriptional and translational processes that produce mRNA transcripts and proteins, respectively. Other examples include molecules (e.g. mRNA and protein) degradation and telomere length shortening processes. In fact, the process of multi-step reactions also exists in other areas such as organic chemistry and biophysical chemistry (Branz, 1996; Qin and Li, 2004). Therefore the major aim of this research work is to design simplified models to accurately characterize biological systems with multi-step reactions.

A widely used approach to simplify multi-step chemical reactions in the literature is to use one-step reaction. For example, the degradation process of mRNA or protein has been modelled by a first order reaction. However, since the one-step reaction cannot provide consistent description of the multi-step reactions, chemical reactions with time delay have been designed recently to model the multi-step chemical events or slow reactions more accurately (Monk, 2003; Zhu *et al.*, 2007; Ma *et al.*, 2005; Burrage *et al.*, 2007). Another important factor is noise in biological networks that may influence the system dynamics substantially. The deterministic modelling methods, which approximate molecular numbers using continuous concentrations (Kaern *et al.*, 2005; Wilkinson, 2009), may not be appropriate to describe systems that contain species with small population numbers. To model stochastic systems more accurately, there are a few other ways. For example, we can use discrete Markov processes where the density of states of a well-stirred chemical reaction system at each time point can be represented by

the chemical master equation (CME) (McQuarrie, 1967; Gillespie, 1992). One of the most well-known methods is called Stochastic Simulation Algorithm (SSA), which is a statistically exact method for simulating trajectories of the CME as the system evolves in time (Gillespie, 1977).

Furthermore, to deal with the intrinsic noise in reactions with time delay, the delay stochastic simulation algorithm (DSSA) was designed by introducing time delay into the SSA (Bratsun *et al.*, 2005; Barrio *et al.*, 2006). Unlike the SSA, which assumes that biochemical reactions are instantaneous and independent, the DSSA characterizes chemical systems that contain both fast and slow reactions. This delayed modelling approach has been applied to many physical and biological systems (Barrio *et al.*, 2006). The DSSA was also extended to describe chemical events that have multiple delays or stochastic delay that follows a given probabilistic distribution (Roussel and Zhu, 2006; Tian *et al.*, 2007a). In recent years, the DSSA has been widely used to simulate the dynamics of genetic regulatory networks and cell signalling pathways (Zhu *et al.*, 2007; Schlicht and Winkler, 2008; Agrawal *et al.*, 2009; Marquez-Lago *et al.*, 2010; Marquez-Lago and Stelling, 2010). In addition, a number of effective simulation methods have been proposed to reduce the huge computing load of the DSSA (Leier *et al.*, 2008; Pahle, 2009; Bayati *et al.*, 2009; Gillespie, 2007). Recently the work done by Mier-y Terán-Romero *et al.* (2010) opened some new aspects for the application of time delays in biological systems. Time delay may not be a constant that was assumed before. Other modelling techniques proposed recently include the slow-scale linear noise approximation and stochastic quasi-steady-state assumption (Thomas *et al.*, 2012; Srivastava *et al.*, 2011). Most recently a new modelling approach has been proposed to simulate chemical reaction systems with memory reactions (Tian, 2013).

The degradation process of mRNA molecules is an important step in the regulation of gene expression, which also represents a typical system with multi-step reactions (Mitchell and Tollervey, 2001). Although the mechanisms of mRNA

degradation have been studied extensively during the last ten years, there are still a number of open problems with respect to the function of enzymes, structure of pathways and role of P-bodies, etc. in the regulation of mRNA degradation (Garneau *et al.*, 2007; Shyu *et al.*, 2008; van Hoof and Parker, 2002). A major step in the quantitative study of mRNA degradation was the development of mathematical models based on the detailed chemical processes. A linear multi-component model was designed to investigate the nonsense-mediated decay of mRNA molecules in yeast (Cao and Parker, 2001, 2003). This deterministic model for mRNA degradation process consists of 23 first-order reactions that describe transcription, translocation, poly(A) shortening, decapping and digestion process. Computer simulations suggested that the widely used concept of half-life underestimated the averaged life-span of mRNA molecules; however, it is still a major factor that determines the life-span of different steps in the degradation pathway. In addition, robustness analysis showed that the change of degradation rate constant led to large variations of mRNA copy numbers. To interpret the complexity of mRNA degradation in a simpler manner, we proposed a multi-step reaction model using a chain of 11 chemical reactions, which gave very good approximation to the detailed one (Tian, 2014).

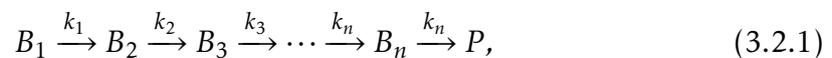
Chemical reactions with time delay has been used to further simplify mathematical models of mRNA degradation. Here time delay represents the time required in the multi-step reactions except the first reaction (Tian, 2014). This simplified model was also extended to using stochastic time delay. However, numerical results showed that these first-order reaction models with delay did not give good approximation to the detailed degradation process (Tian, 2014). Instead of using time delay to represent the missing intermediate reactions in the multi-step reaction, we recently proposed a new modelling approach by introducing a novel concept, namely the length of a molecule indicating its location in the multi-step reactions. Deterministic models using ordinary differential equations have been

used to find the optimal value in a non-linear probability function (Wu *et al.*, 2012). However, it is still a challenge to apply this concept to stochastic models that are much more important than deterministic models for chemical reaction systems. Thus this work further validates the proposed model using stochastic simulations. We first introduce a new stochastic modelling method with two variables for describing chemical events with multi-step reactions, and then propose a stochastic simulation algorithm to numerically calculate the probability of the firing of the last reaction in the multi-step events. The efficiency and accuracy of the proposed method are examined by studying the mRNA degradation process of gene PRL30 based on experimental data.

3.2 Results and Discussion

3.2.1 A new two-variable model

The starting-point of this research work is the chemical events with multi-step reactions. Using the notation proposed in Zhou and Zhuang (2007), we consider the following chemical reactions



where B_i are molecular species and k_i are rate constants. It is assumed that each molecule in the system will eventually turn to the product P or degrade if $P = ()$. During this process, each molecule will pass through a number of states B_1, B_2, \dots, B_n via the multi-step reactions.

When the number of reaction step n is large, we need to design a smaller scale model to simplify the multi-step reactions. We first consider the total number of

molecules in the system, defined by

$$X = \sum_{i=1}^n [B_i]. \quad (3.2.2)$$

Here we introduce a new concept to describe the system state. The number of reactions for a molecule to reach the product P is termed as the length of that molecule. Thus the length of molecule B_i is $(n - i + 1)$ and the total length of the molecules in the system is

$$L = \sum_{i=1}^n (n - i + 1)[B_i]. \quad (3.2.3)$$

According to the total molecule number, chemical reactions in the system can be classified into two groups. If one of the first $(n - 1)$ step reactions occurs, namely $B_i \xrightarrow{k_i} B_{i+1}$, the total number of molecules X is unchanged but the total length L is decreased by one,

$$(X, L) \rightarrow (X, L - 1). \quad (3.2.4)$$

However, if the last reaction $B_n \xrightarrow{k_n} P$ fires, both the total number and total length will decrease by one,

$$(X, L) \rightarrow (X - 1, L - 1). \quad (3.2.5)$$

In this work we use reactions (3.2.4) and (3.2.5) to design the two-variable reaction model.

The key question now is how to determine whether reaction (3.2.4) or (3.2.5) will fire if one of the reactions in the multi-step process (3.2.1) happens. We denote the probability for the degradation of one molecule, namely the firing of reaction (3.2.5), as $f(X, L, n)$, and then the corresponding probability for reaction (3.2.4)

as $1 - f(X, L, n)$. It is clear that, when all molecules are of full length ($X = nL$), the probability of f is zero; while when $X = L$, the probability is one. For the molecules with other lengths, we developed an algorithm, namely Algorithm I in the Method section, to calculate the probability of molecule degradation. With the help of this algorithm, we numerically calculated the exact probability $f(X, L, n)$ using $n = 8$ and $X = 15$ as an example. The probability is represented in Fig. 3.1 as the solid line.

Next we find an appropriate probability function to approximate the calculated curve in Fig. 3.1. Note that the total length L of X molecules satisfies $X \leq L \leq nX$. When $L = X$, all molecules have length 1, the probability of firing of the last step reaction is 1, i.e. $f(X, L, n) = 1$; when $L = nX$, all molecules have length n , there is no chance for the final reaction to occur in the next step, i.e. $f(X, L, n) = 0$. Therefore we suggested a probability function to approximate the curve in Fig. 3.1 in the following format:

$$f(L, X, n) = 1 - \frac{L - X}{X(n - 1)}. \quad (3.2.6)$$

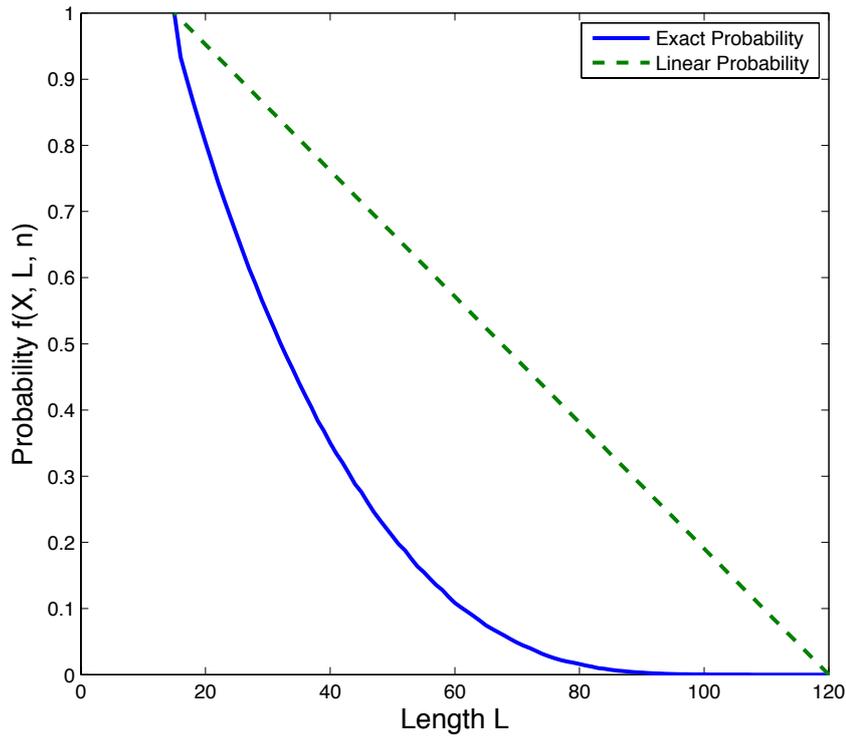


Figure 3.1: The probability for the firing of the last reaction ($B_n \xrightarrow{k_n} P$).

The approximated probability through the proposed function (3.2.6) is plotted as the straight dashed line in Fig. 3.1. It shows that the approximated values are not close to the exact probability values, and the exact probability curve is in a quadratic-like form. Hence, instead of using a linear probability f in terms of X , L and n (3.2.6), we introduced another parameter q into this approximation, and proposed the following two expressions for the probability function f in terms of L , X , n , and q . One candidate is

$$\text{Type I:} \quad f(X, L, n, q) = 1 - \left(\frac{L - X}{X(n - 1)} \right)^q, \quad (3.2.7)$$

and the alternative expression is

$$\text{Type II:} \quad f(X, L, n, q) = \left(1 - \frac{L - X}{X(n - 1)} \right)^q. \quad (3.2.8)$$

3.2.2 Determination of probability function

The major work of this research is to select a probability function from (3.2.7) and (3.2.8) and also search the optimal value of parameter q in the probability function. Using Algorithm I in the Method section, we first calculated the probability $f(X, L, n, q)$ with different values of the total molecule number X ($X = 3 \sim 20$), different numbers of reaction step n ($n = 3 \sim 20$) and various values of the total length L ($L = X \sim nX$). The calculated probability was used as the exact value to search the optimal q in the proposed probability functions. To select a better probability function, we used both type I and II functions to calculate the probability $f(X, L, n, q)$ using the same initial condition ($n = 8$ and $X = 15$) but different values of q ($q = 0.01 \sim 15$) in a step size of 0.01. By searching for a small difference between the exact probability values and those obtained from approximated functions with different q and considering absolute errors only, the optimal values of q for two approximations were achieved, which are 0.27 and 3.91 respectively in this particular case. The exact and approximated probability values are shown in Fig. 3.2. We found that the type II approximation is closer to the exact probabilities than the type I approximation. Then we only used the probability function (3.2.8) for the following studies.

To establish a more general formula for defining q under different conditions of X , L , n , we extended the simulations to various initial conditions of n, X both varying from 3 to 20 together with different values of q . The optimal q values acquired under these conditions are illustrated by Fig. 3.3 (A) and (B). Fig. 3.3 (A) shows that when n increases, the optimal value of q increases for a fixed X value; while Fig. 3.3 (B) indicates that there is no significant variation for the optimal q when X increases for a fixed n value. Therefore, we calculated the averaged optimal q values under various value of n for each given X . A plot of this averaged optimal \bar{q} against n is shown in Fig. 3.3 (C). We suggested a linear relationship between n

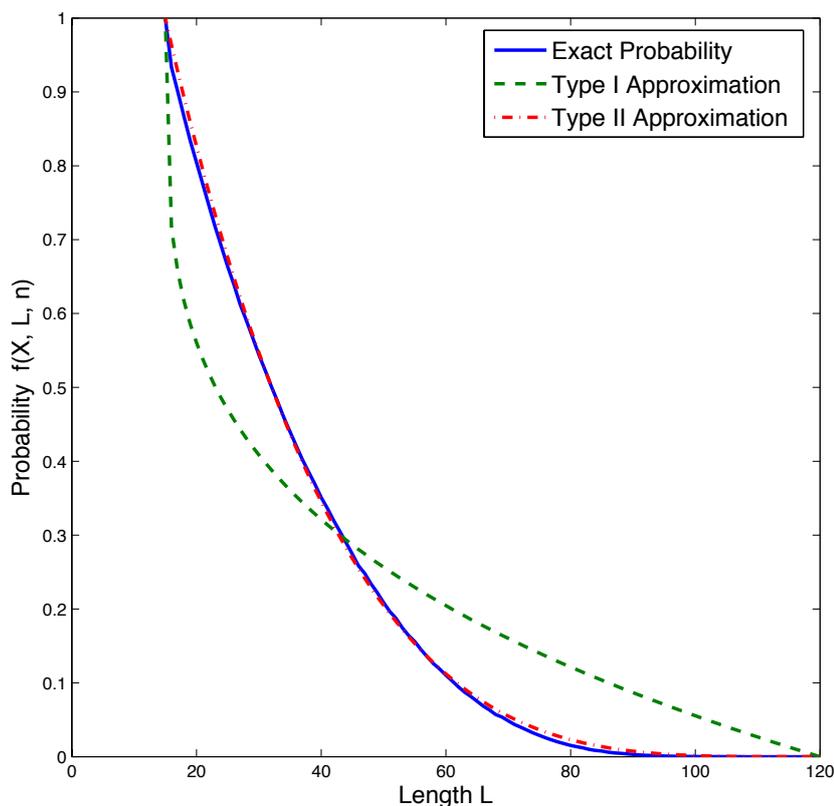


Figure 3.2: Simulated exact probabilities and two approximated probabilities for the firing of the last reaction.

and q . A linear regression analysis suggested that this relationship is

$$\bar{q} = 0.3146n + 1.3615. \quad (3.2.9)$$

We have developed deterministic models of ordinary differential equations (ODEs) based on the multi-step reactions (3.2.1) and the two-variable model (3.2.4, 3.2.5) (Wu *et al.*, 2012). Simulation results of the deterministic models gave some similar patterns such that the optimal value of q increases when the number of chemical reactions n increases. The established relationship between the optimal value of q

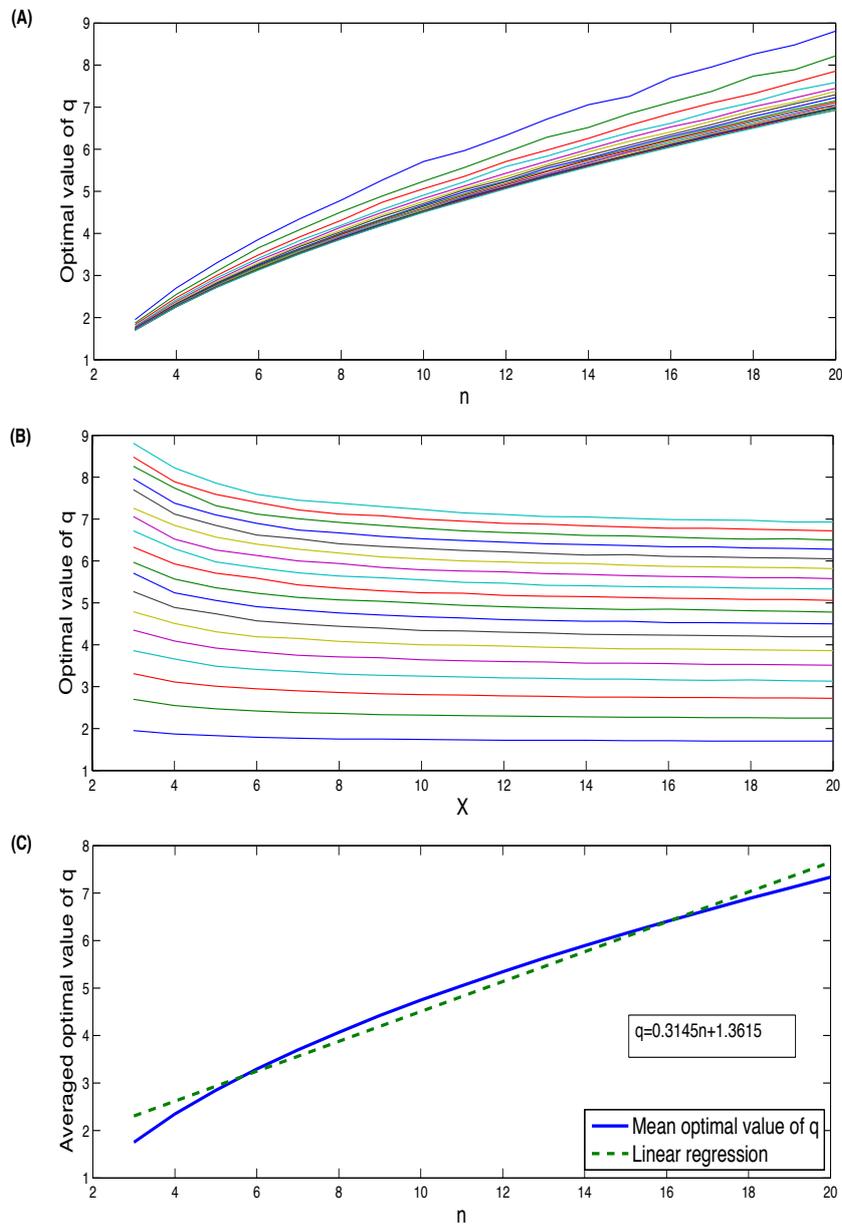


Figure 3.3: Simulation results from probability approach: (A) the optimal values of q with different n ; (B) the optimal values of q with different X ; (C) the averaged optimal values of q against n .

and related model parameters, which is also shown in Fig. 3.4, is given by

$$\bar{q} = 0.4570n + 0.8567. \quad (3.2.10)$$

The above equation is slightly different from expression (3.2.9). For example, when $n = 5$, the averaged value of optimal q is found to be 2.8494 using probability simulation while it is 3.06 from the ODE simulation. The possible reason of the difference is that the ODE model is not the best approach for describing chemical reaction systems with molecules of small copy numbers and some model errors may arise from the ODE simulations. A combination of regression analyses is shown in Fig. 3.4.

Even with the different formulations for the optimal q values, we still find the ODE method confirms the conclusion derived from stochastic simulations. Based on both stochastic and deterministic simulations, we found that the value of q is associated with the number of reaction step n , but not connected to the total molecule number X . Our results also suggested that, when the value of q approaches the optimal one, simulation error of the two-variable model using q is very close to that using the optimal q value. Using the function derived from stochastic simulations, the probability function of molecule degradation is given by

$$f(L, X, n) = \left(1 - \frac{L - X}{X(n - 1)}\right)^{0.3146n + 1.3615}. \quad (3.2.11)$$

Using this probability function, we designed an algorithm, namely Algorithm II in the Method section, to simulate the two-variable reaction model based on the SSA.

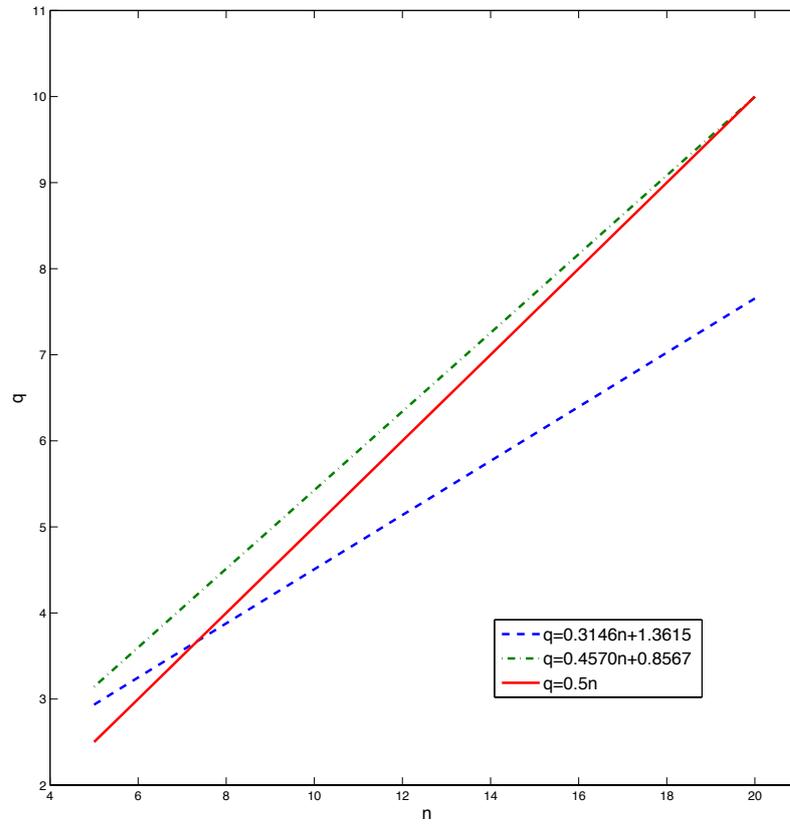


Figure 3.4: Relationship between n and q : Dashed-line: estimated relationship from stochastic simulations; dash-dot-line: relationship derived from the ODE model; Solid-line: $q = 0.5n$.

3.2.3 mRNA decay dynamics: a case study for gene *RPL30*

In this section, we apply the established theory in the previous section to study the dynamics of mRNA degradation. Here we use gene ribosomal protein L30 (*RPL30*) as the test system with a dataset generated from experiments. In these experiments, two constructs of *RPL30* were used to demonstrate the decay kinetics of the mRNA transcripts (Bregman *et al.*, 2011). The first construct (“construct A”) contains the *ACT1* UAS (upstream activating sequence), and the other (“construct B”) contains the *RPL30* UAS. The mRNA molecule decay dynamics was monitored after blocking transcription by using drug 1,10-phenanthroline (Bregman *et al.*,

2011). Thus we assumed that there was no further transcription during the monitoring process. The decay dynamics was normalized by the *RPL30* transcript level at time zero (namely before adding the drug), which was set to 100%. Using the endogenous *RPL30* mRNA levels obtained from the two constructs (Bregman *et al.*, 2011), we first used the one-step differential equation model

$$\frac{dX}{dt} = -kX \quad (3.2.12)$$

to simulate the decay dynamics (Trcek *et al.*, 2011).

Fig. 3.5 (A) and (B) show that the one-step model failed to describe the dynamics of the first 25 minutes accurately. The simulated mRNA levels are always smaller than the experimental observations.

To model mRNA degradation, Cao and Parker (2001) proposed a multi-component model that includes mRNA transcript synthesis, mRNA translocation, poly(A)-shortening process, and terminal deadenylation. We have proposed a simplified model by putting a number of terminal deadenylation reactions into a single one (Tian, 2014). This simple model is a typical multi-step reaction process. In this model, mRNA transcript is synthesized by a zero-order reaction S_1 , then mRNA molecules translocate from the nucleus to cytosol via reaction S_2 . The mRNA molecules in the cytosol produce proteins by the translational process, and in the meantime, the length of mRNA begins to decrease via a number of poly(A)-shortening reactions S_3, \dots, S_9 . The final reaction in this process is the further exonucleolytic degradation S_{10} , which is regarded as the degradation reaction in this work, since the fragment product (*FG*) has no function to produce protein molecules.

Based on the reactions listed in Table 3.1 and rate constants, the propensity functions of these reactions are listed below.

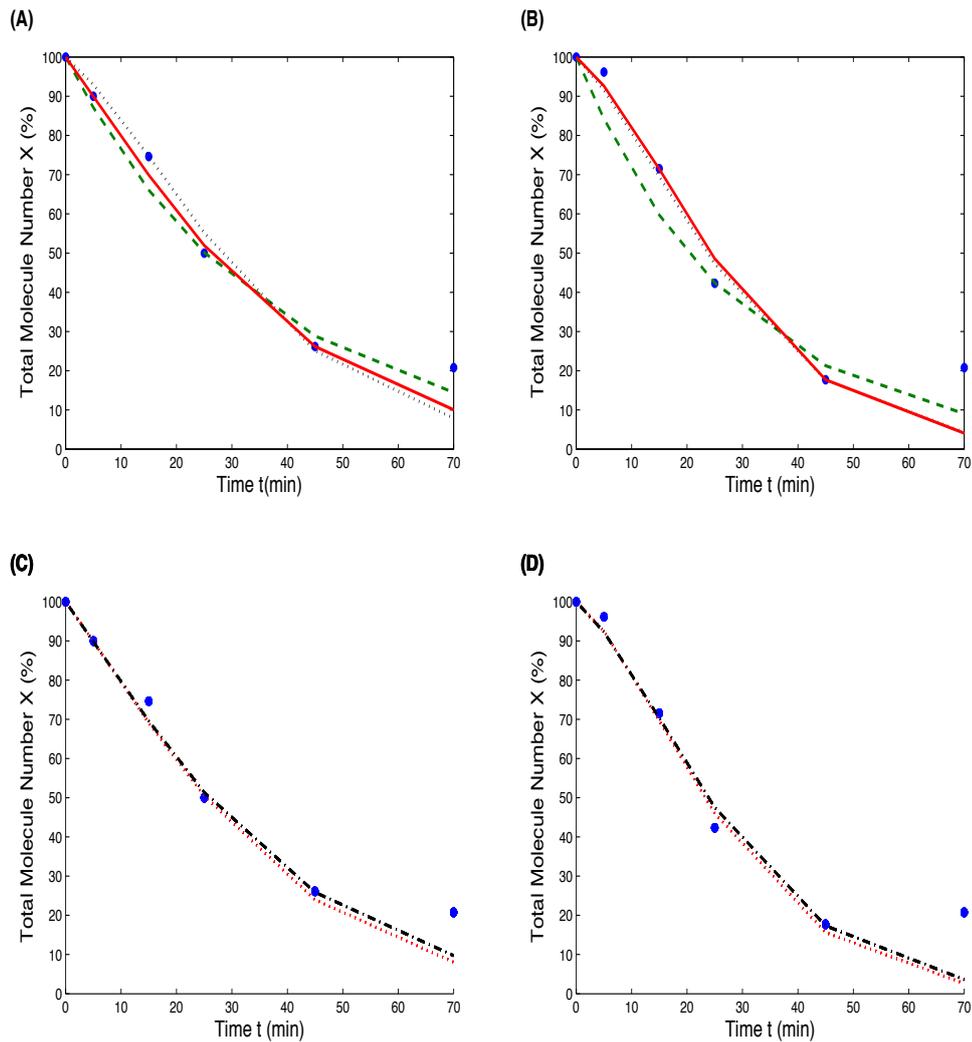


Figure 3.5: Simulated mRNA degradation dynamics using the estimated model parameters: (A) Deterministic simulations for mRNA numbers from the ACT1 construct (green dash-line: the one-step model ($k = 0.0276$), red solid-line: the two-variable model with the optimal initial length ($k = 0.112$, $L = 371$), black dot-line: the two-variable model with the averaged initial length $L = \frac{nX}{2}$, blue dots: experimental data); (B) Deterministic simulations for mRNA numbers from the RPL30 construct (green dash-line: the one-step model ($k = 0.0343$), red solid-line: the two-variable model with the optimal initial length ($k = 0.167$, $L = 473$), black dot-line: the two-variable model with the averaged initial length $L = \frac{nX}{2}$ ($k = 0.161$), blue dots: experimental data); (C) Stochastic simulations of the two-variable model for the ACT1 construct (red dot-line: initial $X_0 = 5$, $k = 0.115$, $L = 19$, black dash-dot-line: initial $X_0 = 10$, $k = 0.111$, $L = 37$, blue dots: experimental data); (D) Stochastic simulations of the two-variable model for the RPL30 construct (red dot-line: initial $X_0 = 5$, $k = 0.171$, $L = 24$, black dash-dot-line: initial $X_0 = 10$, $k = 0.166$, $L = 47$, blue dots: experimental data).

Table 3.1: Reactions, kinetic rates and propensity functions of the simplified stochastic model. The rate constants s_i are in the unit of 1/sec.

	Reaction	Rate constant s_i	Propensity function
S_1	$DNA \rightarrow A$	s_1	$a_1 = s_1$
S_2	$A \rightarrow B$	s_2	$a_2 = s_2 \cdot [A]$
S_3	$B \rightarrow BC1$	s_3	$a_3 = s_3 \cdot [B]$
S_4	$BC1 \rightarrow BC2$	s_4	$a_4 = s_4 \cdot [BC1]$
S_5	$BC2 \rightarrow BC3$	s_5	$a_5 = s_5 \cdot [BC2]$
S_6	$BC3 \rightarrow BC4$	s_6	$a_6 = s_6 \cdot [BC3]$
S_7	$BC4 \rightarrow BC5$	s_7	$a_7 = s_7 \cdot [BC4]$
S_8	$BC5 \rightarrow BC6$	s_8	$a_8 = s_8 \cdot [BC5]$
S_9	$BC6 \rightarrow BC7$	s_9	$a_9 = s_9 \cdot [BC6]$
S_{10}	$BC7 \rightarrow FG$	s_{10}	$a_{10} = s_{10} \cdot [BC7]$

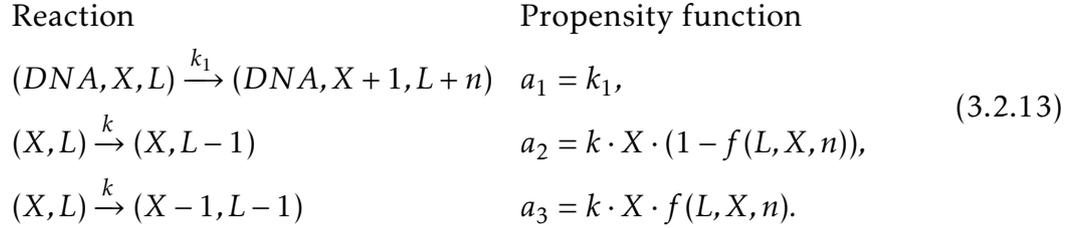
Following the experimental conditions, it is assumed that $s_1 = 0$. For simplicity, it is assumed that $s_2 = \dots = s_{10}$. When using the two-variable model to study the mRNA degradation process, we write the total copy number X and total length of mRNA molecules L as

$$X = [A] + [B] + [BC1] + \dots + [BC7],$$

$$L = 9[A] + 8[B] + 7[BC1] + \dots + [BC7].$$

Here we put the mRNA synthesis as a separate reaction. Then the remaining nine reactions ($n = 9$) form a chemical event of multi-step reactions. The dynamics

of variables $\mathbf{X} = (L, X)$ is described by the following reactions together with the corresponding propensity functions



Using the assumption ($s_2 = \dots = s_{10}$), the rate constant k (3.2.13) is the harmonic mean of rate constants (s_2, \dots, s_{10}), given by

$$k = \frac{n}{\sum_{i=2}^{10} \frac{1}{s_i}} = s_i. \quad (3.2.14)$$

Next we used the proposed two-variable model to give more accurate simulations. We first estimated the degradation rate constant k and optimal initial total length of transcripts. We have also estimated the degradation rate constant k by assuming that the total initial length is a half of the maximal total length ($L = nX/2$), which is termed as the averaged total length. To reduce the computing time, we first estimated parameters in the ODE model (3.2.12) using different initial transcript numbers ($X_0 = 5, 10, 20, \dots, 100$). Table 3.2 suggests that the variation between the estimate rate constant k was very small for different initial mRNA numbers. Similar observation is applied to the ratio of the optimal initial total length to the maximal total initial length, namely $L_0/(nX)$, for the tests with different initial mRNA numbers. Thus our results suggested that the estimated model parameters are independent to the initial mRNA copy numbers.

Table 3.2: Estimated parameters for the stochastic model of RPL30 and ACT1 mRNA degradations (Ratio= L_0/nX).

X_0	ACT1 construct			RPL30 construct		
	Rate k	L_0	Ratio	Rate k	L_0	Ratio
$m = 5$	0.1150	19	0.4222	0.1710	24	0.5333
$m = 10$	0.1110	37	0.4111	0.1660	47	0.5222
$m = 20$	0.1130	75	0.4167	0.1680	95	0.5278
$m = 30$	0.1130	112	0.4148	0.1670	142	0.5259
$m = 40$	0.1120	149	0.4139	0.1680	190	0.5278
$m = 50$	0.1120	186	0.4133	0.1670	237	0.5267

Using the estimated model parameters of the case $X_0 = 100$, simulation results for the two constructs in Fig. 3.5 (A) and (B) show that the two-variable model provides more accurate description of the mRNA degradation dynamics than the one-step model, in particularly for that in the first 25 minutes. For the ACT1 construct in Fig. 3.5 (A), the optimal length number with ratio 0.412 gave more accurate simulation than the averaged length number. However, in Fig. 5 (B) for the RPL30 transcript, the difference between the simulations using two different length numbers is small. In this case, the optimal ratio is 0.525, which is very close to 0.5.

To further examine the accuracy of the two-variable model, we used the stochastic model to simulate the mRNA dynamics using different initial transcript numbers. For each initial mRNA number, we generated 10,000 simulations and then calculated the averaged mRNA numbers of all stochastic simulations. For both constructs in Fig. 3.5 (C) and (D), our results show that there is small difference

between the simulations using $X_0 = 5$ and $X_0 = 10$. However, there is not any significant difference between simulations when the initial mRNA number is larger than 10.

Finally we provided a few stochastic simulations for mRNA degradation dynamics for the construct *ACT1* in single cells. The rate constants of the detailed model were derived from the two-variable model using the relationship (3.2.14); and the initial molecular numbers were randomly selected while the length of the initial molecules matches the length in the two-variable model. When the mRNA synthesis rate is $s_1 = 0$, Fig. 3.6 shows that the molecular numbers and lengths approach to zero at the time point around 100. In addition, compared with the simulations of the detailed model in Fig. 3.6 (A) and (C), the two-variable model generates simulations with more fluctuations in Fig. 3.6 (B) and (D). After the time point 100, more simulations of the two-variable model still have non-zero molecular numbers.

3.3 Conclusion

This work represents an attempt to use simplified mathematical models to describe complex biological systems. Concentrating on the chemical events of multi-step reactions, we proposed a new concept (e.g. the length of a molecule) as an additional measure to characterize system dynamics. The length of a molecule is defined as the location of a molecule in the multi-step reactions. Using the total molecule number and total length of molecules, we proposed a two-variable model to reduce the complexity of the multi-step reactions. The major contribution of this work is to design a non-linear function that represents the probability of the firing of the last reaction in the multi-step reactions. To calibrate this probability function, we proposed a stochastic simulation method to calculate the probabilities of various system states. Numerical results suggested that this

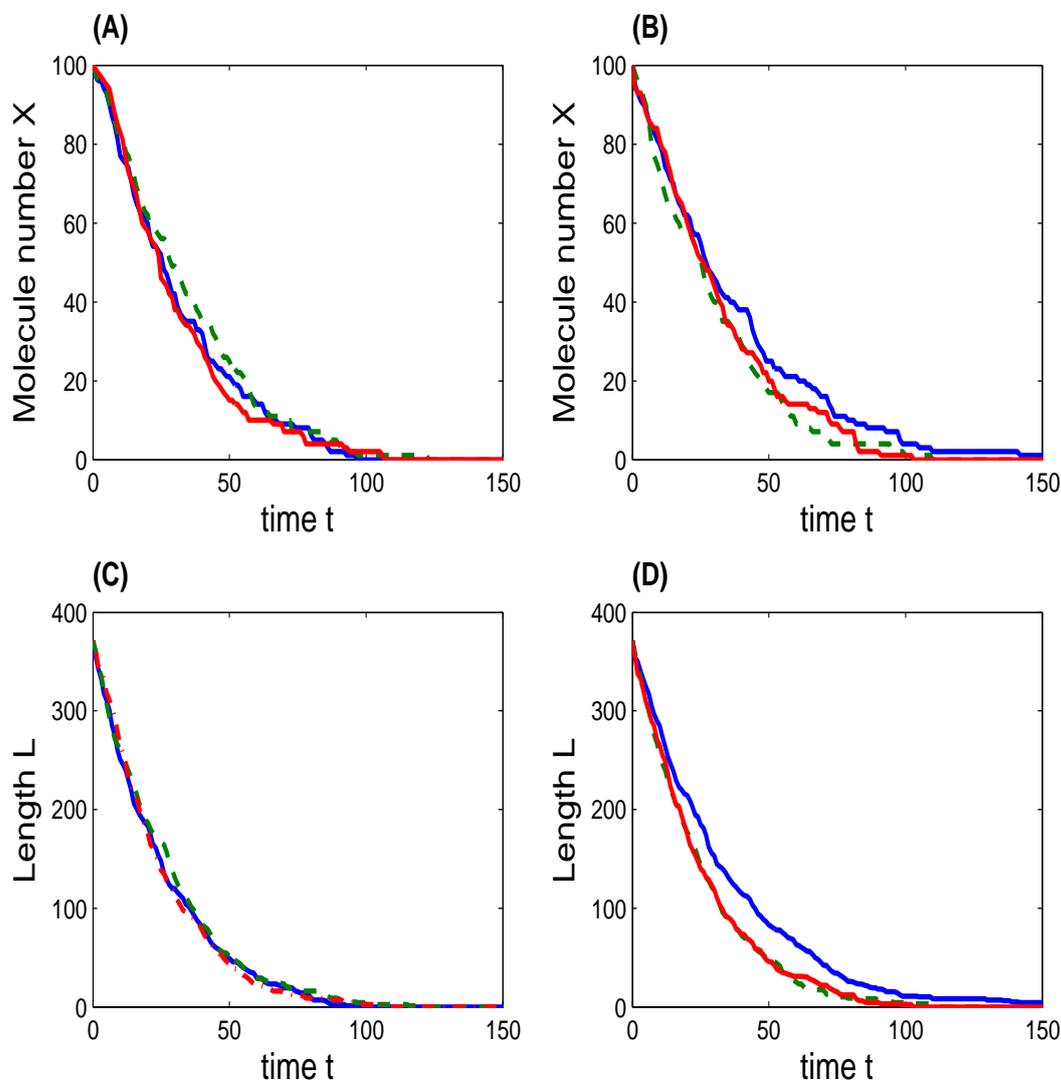


Figure 3.6: mRNA degradation dynamics of gene RPL30 construct ACT1 in single cells: (A,C) three simulations of X and L values over t for the detailed multi-step reaction model. (B,D) three simulations of X and L values over t for the two-variable model. For the detailed multi-step model, rate constants are $s_1 = 0$, $s_2 = \dots = s_{10} = 0.112$, and initial molecular numbers are $([A], [B], [BC1], \dots, [BC7]) = [4, 3, 3, 5, 15, 20, 20, 15, 15]$. For the two-variable model, rate constant is $k = 0.112$, initial conditions $(X_0, L_0) = (100, 371)$.

probability is dependent on the number of reaction steps but independent of the total molecule number, which suggested that we were able to design a simplified model based on the network structure. Then our proposed two-variable model was applied to simulate the dynamics of mRNA degradation using experimentally observed data. Numerical results suggested that the length of molecules, which is approximately a half of the maximal length initially, played an important role in realizing experimental data. The potential future work includes the application of the two-variable model to other multi-step reaction systems such as gene expression and telomere length regulation. In addition, the refinement of the two-variable model, such as the accuracy of the probability function, would also be very interesting.

3.4 Methods

3.4.1 Simulation algorithm for the probability function

To find the probability for the firing of the last reaction in the multi-step reactions (3.2.1), we first designed a Monte-Carlo method to numerically calculate the probability function $f(X, L, n)$ based on the given X and n . By the law of total probability, the formation of probability that the final reaction occurs given by any L and X is defined as following:

$$f(X, L, n) = \sum_{j=0}^X P(R_n | B_n = j) \cdot P(B_n = j | L, X),$$

where R_n represents the occurrence of last reaction $B_n \xrightarrow{k_n} P$. Based on the total molecule number X , we calculate the probability

$$P(R_n | B_n = j) = \frac{j}{X}.$$

The major part of this algorithm is to find frequency of the event $B_n = j$ based on the given length L and total molecule number X , which is explained as the following algorithm I.

Algorithm I

1. Set the total number of molecule X , number of reactions n , and initial full length $L_0 = nX$.
2. Based on the following 10,000 Monte-Carlo simulations, calculate the frequency $freq(B_n = j|L, X)$ for X molecules with total length L having j molecules, where $j = 0, 1, \dots, X$ and each of the molecule with length 1:
 - (a) Consider X molecules with full length. Denote the length of the i -th molecule as l_i with $i = 0, 1, \dots, X$.
 - (b) Use a random number $r \sim U(0, 1)$ to select one molecule, with index j . If the length of that molecule $l_j > 1$, reduce its length by 1, namely $l_j = l_j - 1$; if $l_j = 1$, then repeat this step until finding a molecule with length greater than 1.
 - (c) Repeat step (b) for $(L_0 - L)$ times to get a set of molecules with total length L .
 - (d) Count the number of molecules in this set with length 1, denote as i , then update

$$freq(B_n = i|L, X) = freq(B_n = i|L, X) + 1.$$

- (e) Repeat steps (a) ~ (d) for 10,000 times.

3. The probability for the last reaction firing is obtained by

$$f(X, L, n) = \sum_{j=1}^X \frac{freq(B_n = j|L, X)}{10000} \times \frac{j}{X}.$$

3.4.2 Ordinary differential equation model

The most widely used approach to study chemical reaction systems is deterministic model using ordinary differential equations. The approach is valid if the copy numbers of chemical species in the system are large. To confirm the probability function $f(X, L, n)$ derived from stochastic simulations, we designed a deterministic model of ODEs for the multi-step chemical reaction system (3.2.1), given by

$$\begin{aligned}\frac{dB_1}{dt} &= -k_1 B_1, \\ \frac{dB_2}{dt} &= k_1 B_1 - k_2 B_2, \\ &\vdots \\ \frac{dB_n}{dt} &= k_{n-1} B_{n-1} - k_n B_n, \\ \frac{dP}{dt} &= k_n B_n.\end{aligned}\tag{3.4.1}$$

Using the total molecule number $X (= B_1 + \dots + B_n)$ and the total length of the molecules $L (= B_n + 2B_{n-1} + \dots + nB_1)$, we have a simplified model of the above ODE system

$$\begin{aligned}\frac{dX}{dt} &= -k B_n, \\ \frac{dL}{dt} &= -k X,\end{aligned}\tag{3.4.2}$$

where k is the harmonic mean of the rate constants k_1, \dots, k_n (3.4.4), and $k B_n$ represent the probability of molecule degradation, which is represented by the probability function $f(X, L, n)$. Using the notations of stochastic simulation, the

ODE model with the length of molecules is given by

$$\begin{aligned}\frac{dX}{dt} &= -kX \left(1 - \frac{L-X}{X(n-1)}\right)^q, \\ \frac{dL}{dt} &= -kX.\end{aligned}\tag{3.4.3}$$

For a given initial condition $B_i(0)$, we obtained the analytical solution of the detailed system (3.4.1) and then solved the two-variable model (3.4.3) numerically using a stiff ODE solver *ode23s* in MATLAB. We tested the solution of the two-variable model with different values of q based on different system conditions ranging from $n = [5 \ 10 \ 15]$ as well as $X = [5 \ 10 \ 50 \ 100 \ 200 \ 500]$. For each system condition, we selected the optimal value of q with which the two-variable model (3.4.3) generates simulation that is very close to that of the detailed ODE model (3.4.1). Finally we find the relationship between the value of q and system condition (X, L, n) by using a regression method (Wu *et al.*, 2012).

3.4.3 An algorithm for simulating systems including two-variable model

The SSA is a general framework for simulating biochemical reaction systems. Now we propose an algorithm to incorporate the two-variable model into the SSA. It is assumed that a chemical reaction system is a well-stirred mixture at constant temperature in a fixed volume Ω . This mixture consists of N molecular species $\{S_1, \dots, S_N\}$ that chemically interact through M reaction channels $\{R_1, \dots, R_M\}$. The dynamic state of this system is denoted as $\mathbf{x}(t) \equiv (x_1(t), \dots, x_N(t))^T$, where $x_i(t)$ is the molecular number of species S_i at time t . For each reaction channel $R_j (j = 1, \dots, M)$, a propensity function $a_j(\mathbf{x})$ is defined by a given state $\mathbf{x}(t) = \mathbf{x}$ and the value of $a_j(\mathbf{x})dt$ represents the probability that one reaction will occur somewhere during the infinitesimal time interval $[t, t + dt)$ (Gillespie, 1977, 2001,

2007). In addition, a state change vector v_j is defined to characterise the change of molecular numbers due to the reaction R_j . The element v_{ij} of v_j represents the change of the copy number of species S_i . The algorithm for simulating chemical reaction systems with two-variable model is given below.

Algorithm II

1. Calculate the values of propensity function $a_j(\mathbf{x})$ based on the system state \mathbf{x} at time t . In particular, for the two-variable reaction with the total molecule number X (3.2.2) and total length L (3.2.3), the propensity function is $a_j = kX$, where k is the harmonic mean of the rate constants (3.2.1), given by

$$k = \frac{n}{\frac{1}{k_1} + \dots + \frac{1}{k_n}}. \quad (3.4.4)$$

Then the sum of propensity function values is

$$a_0(\mathbf{x}) = \sum_{j=1}^M a_j(\mathbf{x}).$$

2. Generate a sample r_1 of the uniformly distributed random variable $U(0,1)$, namely $r_1 \sim U(0,1)$, and determine the time of next reaction

$$\mu = \frac{1}{a_0}(\mathbf{x}) \ln \frac{1}{r_1}.$$

3. Generate another sample r_2 of $U(0,1)$ to determine the index k of the next reaction occurring in $[t, t + \mu]$

$$\sum_{j=1}^{k-1} a_j(\mathbf{x}) < r_2 a_0(\mathbf{x}) \leq \sum_{j=1}^k a_j(\mathbf{x}).$$

4. If the k -th reaction is not a two-variable model, update the state of the system by

$$\mathbf{x}(t + \mu) = \mathbf{x}(t) + v_k.$$

Otherwise generate a sample $r_3 \sim U(0, 1)$ to determine which reaction of the followings will occur,

$$(X, L) = \begin{cases} (X, L - 1) & \text{if } r_3 > f(X, L, n), \\ (X - 1, L - 1) & \text{if } r_3 < f(X, L, n), \end{cases}$$

where $f(X, L, n)$ is the probability of the firing of the last reaction. Then the system is updated.

5. Go back to step 1 if $t + \mu \leq T$, where T is the end time point. Otherwise, the system state at T is $\mathbf{x}(T) = \mathbf{x}(t)$.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 4

Declaration by candidate

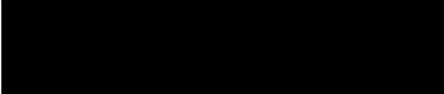
In the case of Chapter 4, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Developed, established and verified the method Wrote programming codes and the article	90%

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Tianhai Tian	Provided helpful guidance and editorial work	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature		Date 29/04/15
Main Supervisor's Signature		Date 29/04/15

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 4

Stochastic Modelling of Regulatory Networks using State-dependent Time Delay

Chapter 4 is based on the article Wu Q, Tian T. 2015. Stochastic modelling of regulatory networks using state-dependent time delay (to be submitted) .

Abstract. *The advances of systems biology have raised a large number of mathematical models for describing the dynamics of molecular regulatory networks. To deal with the growing scale of molecular systems, sophisticated modelling techniques have been designed in recent years to reduce the complexity of mathematical models. Among them, a widely used approach is to use a one-step “slow” reaction or delayed reaction to simplify multi-step reactions. However, recent research results suggest that a delayed reaction with constant time delay is still unable to describe multi-step reactions accurately. To address this challenge, this work proposes a novel approach using state-dependent time delay. We first use the stochastic simulation algorithm to calculate time delay for multi-step reactions exactly, which clearly shows that the value of time delay depends on the dynamics of system states. Then we design an algorithm to calculate the value of time delay precisely. To demonstrate the power of the proposed method, we use two processes of mRNA degradation to investigate the function of time-delay in determining the system dynamics. Simulation results of the first model show that molecules in different stages of degradation all are important to calculate the half-life of mRNA molecules. The second model provides further evidences for the importance of state-dependent time delay in gene expression and mRNA degradation. These results suggest that we may need to reconsider the concept of half-life for measuring the degradation process of molecules.*

Keywords. *Genetic regulation; stochastic modelling; time delay; mRNA degradation.*

References are considered at the end of the thesis.

Chapter 4

Stochastic Modelling of Regulatory Networks using State-dependent Time Delay

4.1 Introduction

This process of gene expression is complex and includes a number of key steps, such as transcription initiation, RNA polymerase elongation, mRNA translocation, and translation. Each step may include a series of detailed chemical reactions. Due to the low copy number of molecules in this process, a gene is activated and inactivated by random association and dissociation transcriptional factors (TFs) and other events. Recent advances in experimental technology have provided the ability to measure and interpret cellular heterogeneity in single cells (Raj and van Oudenaarden, 2009; Spiller *et al.*, 2010; Srivastava *et al.*, 2011). Following the observation of translational bursts (Cai *et al.*, 2006; Ozbudak *et al.*, 2002), single cell studies demonstrate that gene transcription also occur in bursts of multiple transcripts separated by relatively long periods of transcriptional inactivity

(Chubb and Liverpool, 2010). Experimental studies in recent years have shown that gene expression is subject to stochastic fluctuations that lead to considerable differences in the level of expression between genetically identical cells (Kaern *et al.*, 2005). In addition, experimental data suggest that variation in protein levels arises from fluctuations in mRNA levels due to random production and decay of mRNAs or random activation and inactivation of the gene promoter (Kaufmann and van Oudenaarden, 2007).

Stimulated by the pioneer work of stochastic modelling and experimental advances in single cell studies (Arkin *et al.*, 1998), the last ten years have seen an explosion in stochastic modelling of these processes to predict protein fluctuations in terms of the frequencies of probabilistic event (Burgess, 2014; Padovan-Merhar and Raj, 2013). Although the complex biological processes can be modelled by a series of detailed chemical events such as the processes in gene expression, the structure of the model may be too complex to get any insights mathematically. To address this issue, a number of modelling techniques have been proposed to simplify the complexity of mathematical model (Burrage *et al.*, 2004; Bokes *et al.*, 2012). Among them, differential equation with time delay has been used to simplify processes of multi-step reactions (Monk, 2003). To explore the combined effects of time delay and intrinsic noise on gene regulation, delay stochastic simulation algorithm (delay-SSA) (Barrio *et al.*, 2006; Bratsun *et al.*, 2005) has been proposed to simulate discrete chemical kinetic systems. The advances in delayed modelling approaches include mathematical model for the translational process to include spatial effects in gene expression (Marquez-Lago *et al.*, 2010), and the linear-noise approximation for stochastic reaction systems with distributed delays (Brett and Galla, 2013). Other modelling techniques proposed recently include the slow-scale linear noise approximation and stochastic quasi-steady-state assumption (Srivastava *et al.*,

2011; Thomas *et al.*, 2012; Ribeiro, 2010). Recently, we have proposed a two-variable model that with concept of length and stochastic simulation algorithm for memory reactions (Wu *et al.*, 2013b; Tian, 2013).

Currently it is widely assumed that time delay is either a constant or distributed delay with constant mean. For example, a one-step reaction model with constant, exponentially distributed or Erlang distributed delays has been used to realize the mRNA turnover dynamics (Tian, 2014). Our simulation results suggested that time delay may depend on the system state, rather than be a constant value. In fact, the state-dependent time delay has already been used in various research areas such as optimal control and population dynamics (Asher and Sebesta, 1971; Cao *et al.*, 1992). Although these ideas were proposed about 40 years ago, the relationship between the time delay and system state remains uncertain for discrete chemical reaction settings. Recently, a delay model with non-constant time delay has been derived using an analytical method to simplify the translational process of multi-step reactions (Mier-y Terán-Romero *et al.*, 2010). However, more work is needed to address this issue for the widely used multi-step reactions.

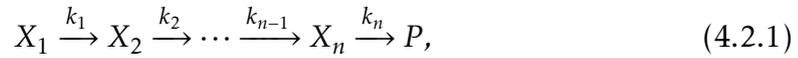
The motivation of this work is to develop a new method using chemical reaction with state-dependent time delay to simplify multi-step reactions accurately. The proposed method will be validated by the degradation process of mRNA molecules, which is a typical multi-step chemical reactions system. This degradation process has drawn much attention from researchers during the last ten years. In traditional experimental studies, a large sample of cells are genetically modified or treated with inhibitors to stop transcription and thus kinetic information of a decaying mRNA species can be obtained (Passos and Parker, 2008). Recently single-cell and single-molecule techniques have advanced our understanding of mRNA turnover. The kinetic behaviour of individual RNA polymerase II (RNAPII) transcribing a gene provides a precise quantification of the contribution of mRNA synthesis to the cellular pool of transcripts (Ardehali and Lis, 2009). The accuracy of decay

measurement varies with the technique used. For example, in budding yeast, half-lives of an individual mRNA species quantified by different approaches may differ by more than 50% (Grigull *et al.*, 2004; Holstege *et al.*, 1998; Wang *et al.*, 2002). Although, a detailed mechanistic model has been designed to describe the degradation process exactly (Cao and Parker, 2001, 2003), it is difficult to derive accurate information of half-life from detailed mechanistic models. Thus precise analysis of decay kinetics is strongly needed to provide more information regarding the half-life of mRNA molecules.

4.2 Methods

4.2.1 Multi-step chemical reaction system

This study considers the following system with a series of chemical reactions:



where X_i represents the i -th state of a molecule and k_i is rate constant. Here “ P ” is the product, which may also be “()” if it is a degradation process. Denote s as the total copy number of molecules in all states, namely $s = \sum_{i=1}^n x_i$, where x_i is the copy number of state X_i . The dynamics of system (4.2.1) can be described by an ordinary differential equation (ODE) model, which is given in the Supplementary Information. For simplicity, it is assumed that $k_1 = k_2 = \dots k_n = k$. Then the exact solution of this ODE model is derived in Supplementary Information. In particular, the exact solution of the total molecule number is given by

$$s = s_0 e^{-kt} + (s_0 - x_{n0}) k t e^{-kt} + (s_0 - x_{(n-1)0} - x_{n0}) \frac{(kt)^2}{2!} e^{-kt} + \dots + \frac{x_{10}}{(n-1)!} (kt)^{n-1} e^{-kt}. \quad (4.2.2)$$

where x_{i0} is the initial copy number of X_i and $s_0 = (x_{10} + \dots + x_{n0})$.

4.2.2 State-dependent time delay

We use a reaction with time delay to simplify system (4.2.1), which is described as follows:



Here reaction (4.2.3) is the first reaction of system (4.2.1), while delayed reaction (4.2.4) is a simplification of the process from state X_2 to product P . The time delay in reaction (4.2.4) is the sum of waiting time experiencing $n - 1$ consecutive reactions from state X_2 to product P . Thus the imaginary state G represents any one of the intermediate states X_2, \dots, X_n and thus its molecular number is $y = \sum_{i=2}^n x_i$.

The question we are interested now is, for any given system state with copy number $[x_1, y]$, how to exactly calculate time delay for the next molecule of X_1 turning to product P . It is assumed that the newly created imaginary molecule forming from a deduction of X_1 (namely, copy number from x_1 to $x_1 - 1$) will be manifested to product after the current y imaginary molecules turn to product. Denote τ_1 as the time point when copy number of X_1 decreases from x_1 to $x_1 - 1$; and τ_2 the time point when the total copy number s decreases from $x_1 + y$ to $x_1 - 1$. Then the time delay is

$$\tau = \tau_2 - \tau_1 \quad (4.2.5)$$

Here the value of τ_1 is the waiting time of the first reaction in (4.2.1), which is determined by the stochastic simulation algorithm. In Supplementary information,

we derive that the value of τ_2 as

$$\tau_2 = \begin{cases} \frac{1+x_1 n-x_1}{kx_1} & \text{if } C_1 = 0 \\ -\frac{2n}{k} W\left(-\frac{1}{2n} C_1^{\frac{1}{n}}\right) & \text{if } C_1 \neq 0 \end{cases}, \quad (4.2.6)$$

where $W(x)$ is the Lambert W function, and

$$C = \frac{(1 + C_2 y)n!}{C_1}, \quad (4.2.7)$$

$$C_1 = x_1 + y - \frac{ny}{n-1}. \quad (4.2.8)$$

However, the value of C_2 is dependent on the values of x_1 and y , which will be determined in Section (4.3) by numerical simulations.

4.2.3 SSA with state-dependent time delay

This work proposes the following modelling framework with time delay. We need to simulate a well-stirred mixture of $N(\geq 1)$ molecular species $\{X_1, \dots, X_N\}$ that chemically interact inside some fixed volume Ω at a constant temperature and through M reaction channels $\{R_1, \dots, R_M\}$, which includes M_1 elementary reactions and M_2 delayed reactions ($M = M_1 + M_2$). Here a delayed reaction may be a reaction with constant time delay, distributed delay that follows a distribution, or state-dependent time delay that is simplified from the lumped multi-step chemical reactions (4.2.1). The system state is denoted as $X(t) \equiv \{x_1(t), \dots, x_N(t)\}^T$, where $x_i(t)$ is the copy number of species X_i . For each delayed reaction, we define an imaginary species G_i to represent the intermediate species of that delayed reaction. We also define a stoichiometric vector v_j for non-delayed reactions, as well as consuming and manifesting stoichiometric vectors v_j and u_j for delayed reactions, respectively. For each reaction channel, a propensity function $a_j(X)$ is defined and

$a_j(X)dt$ represents the probability of this reaction will fire inside Ω in the next infinitesimal time interval $[t, t + dt]$. Detailed algorithm is given below.

Algorithm: State-dependent Delay SSA(SD-SSA)

Set initial molecular numbers at $t = 0$, and an empty queue structure L for storing the information of delayed reactions.

- Step 1: Calculate propensity functions $a_j(X)$, $j = 1, \dots, M$ and $a_0(x) = \sum_{j=1}^M a_j(X)$.
- Step 2: Generate a uniform random number $r_1 \in U(0, 1)$ and determine the waiting time of the next reaction

$$\tau_1 = -\ln \frac{r_1}{a_0}$$

- Step 3: Compare δ with the least time δ_{min} in the queue structure L to check whether there is any delayed reactions that are scheduled to finish within $[t, t + \tau_1)$.
- Step 4: IF $\delta_{min} < \tau_1$ (Update the delayed reaction at δ_{min})

$$X(t + \delta_{min}) = X(t) + u_j.$$

ELSE: Determine the index j of next reaction by a uniform random number $r_2 \in U(0, 1)$

$$\sum_{k=1}^{j-1} a_k(X) < r_2 a_0(X) \leq \sum_{k=1}^j a_k(X)$$

and update the system state by

$$X(t + \tau_1) = X(t) + v_j.$$

Then determine time delay for the possible delayed reaction. Use the constant delay for the normal delayed reaction; generate a sample for the distributed delay reaction; and use (4.2.5,4.2.6) to calculate the delay value τ if R_j is a reaction with state-dependent time delay. Then add index j and update time $t + \tau_1 + \tau$ to the queue structure L .

- Step 5: Go to Step 2.

Note that this delayed simulation algorithm is based on the so-called rejection method delay-SSA (Barrio *et al.*, 2006). A more precise algorithm can be considered if we consider the change of propensity functions due to the update of a delay reaction in step 2.

4.3 Results

4.3.1 State-dependent time delay

To demonstrate the dependence of time-delay on system state, we first apply SSA to numerically calculate the value of delay under various initial conditions, which is given in Supplementary Information as Algorithm 1. We first test the case with different values of x_{10} but fix $y_0 (= 0)$. Total molecular number will decrease from $s_0 - i + 1$ to $s_0 - i$ when the i -th delay reaction occurs. Similar to the notations in (4.2.5), we denote τ_{1i} as the point when x_1 decreases from $x_{10} - i + 1$ to $x_{10} - i$ while τ_{2i} as the time when s decreases from $s_0 - i + 1$ to $s_0 - i$. Then the delay time for the i -th molecule is $\tau_i = \tau_{2i} - \tau_{1i}$. Fig. 4.1 (A) gives calculated values of time delay based on different initial conditions (namely $x_{10} = 5, 10, 20, 40$). For each initial condition, the value of time delay increases when the total molecular number decreases, which is due to the small value of propensity functions that leads to large waiting time for chemical reactions. For example, when the initial

molecular number is $x_{10} = 40$, the time delay for the decay of the first molecule is $t = 38.2$, while that for the last molecule is $t = 125.3$. Similarly, if the initial molecular number x_{10} is larger, the delay time for the molecule of the same order is smaller. These results clearly suggest that the value of time delay depends on the value of propensity functions that are determined by the system state.

To further demonstrate the dependence of time delay on imaginary molecules, we also calculated time delay for the decay of the first molecule based on different initial molecule number x_{10} and imaginary molecule number $y_0 (> 0)$. In SSA the value y_0 is transferred to the initial molecule numbers $(x_{20}, x_{30}, \dots, x_{n0})$. It is assumed that the initial values of x_{i0} satisfy $x_{20} \geq x_{30} \geq \dots x_{n0}$ and the difference between these numbers is at most 1. For example, if $n = 5$ and $y_0 = 6$, the initial system state follows as $(x_{10}, x_{20}, \dots, x_{n0}) = (x_{10}, 2, 2, 1, 1)$ for a given initial molecule number x_{10} . Fig. 4.1 (B) suggests that the calculated time delay also depends on the number of imaginary species in the system. In addition, Fig. 4.1 (B) suggests that the change of value y_0 has different impact on the value of time delay. For example, when $n = 5$, the increment of y_0 from 7 to 8 does not lead to much change of time delay. In this case, the variation of initial condition is from $(x_{10}, \dots, x_{50}) = (x_{10}, 2, 2, 2, 1)$ to $(x_{10}, 2, 2, 2, 2)$. However, when the value of y_0 is increased from 8 to 9, the change of time delay is large because the initial state is now $(x_{10}, 3, 2, 2, 2)$.

4.3.2 Formula for calculating time delay

We have derived an expression for calculating time delay based on a given system state. However, an unsolved question is the value of C_2 in (4.2.6) that explicitly includes time t (see Supplementary Information). Here we find an approximation of C_2 through numerical computations. We first calculate the optimal value of C_2 for different values of x_{10} and y_0 using the derived expressions (4.2.5, 4.2.6) to

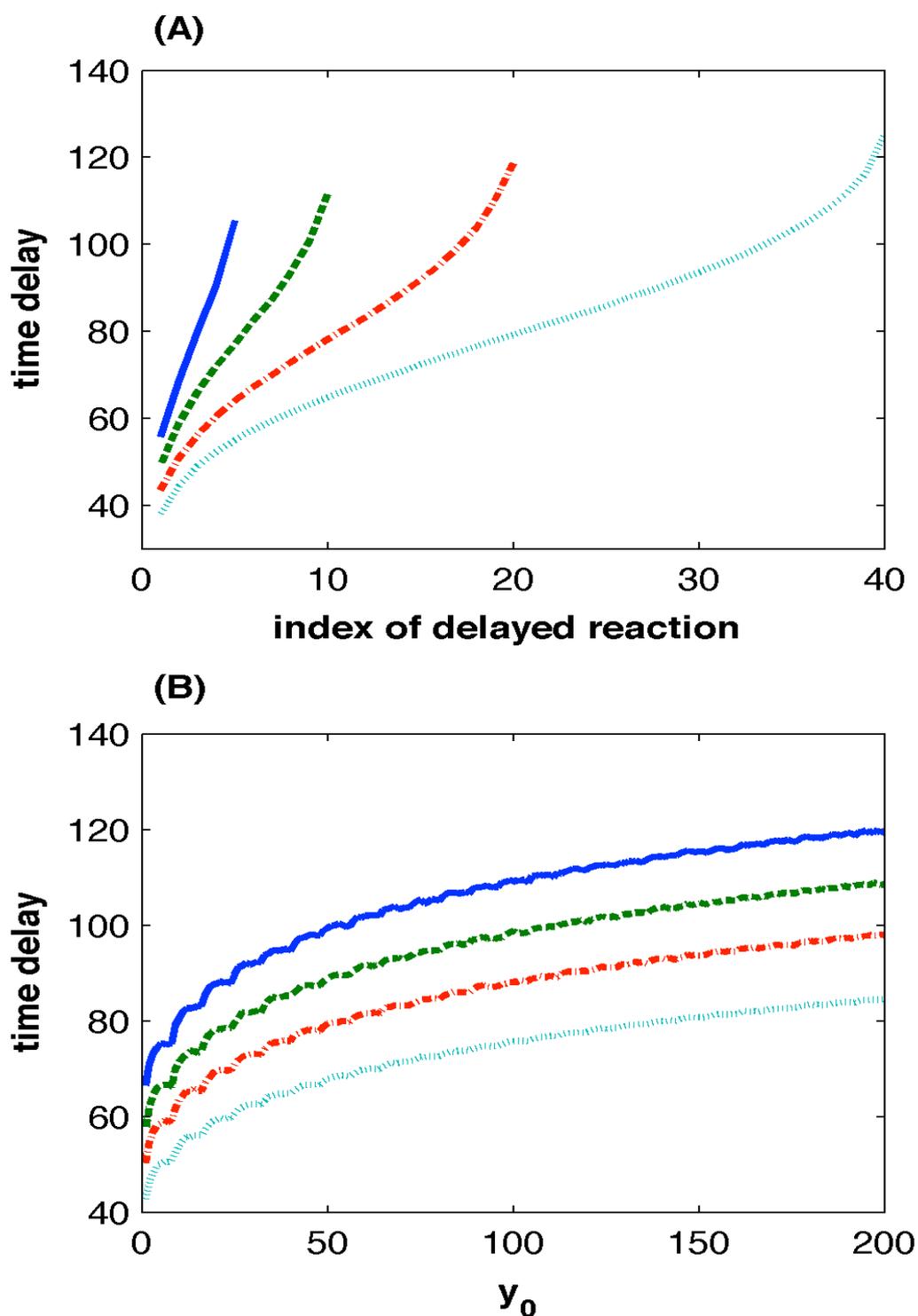


Figure 4.1: Calculated time delay using stochastic simulations of the multi-step reactions process (4.2.1): (A) Time delay for the decay of each molecule based on different initial number x_{10} but null initial imaginary species y_0 . Index i means the delay of the i -th molecule. (B) Time delay for the decay of the first molecule based on different values of x_{10} and y_0 . (Solid-line: $x_{10} = 5$, dash-line: $x_{10} = 10$, dash-dot-line: $x_{10} = 20$, dot-line: $x_{10} = 40$).

match calculated time delays in Fig. 4.1 (B). The optimal values of C_2 in Fig. 4.6 (A) suggest that it is a monotonically decreasing function of y_0 . In addition, the value of C_2 is $-1/y$ when $y_0 = x_1(n-1)$. Thus we assume that

$$C_2 = \frac{\alpha(x_1(n-1) - y)}{\beta + y} - \frac{1}{y} \quad (4.3.1)$$

To determine the values of α and β in (4.3.1), we further estimate these values by matching the determined time delay using expression (4.2.5 ~ 4.3.1) with those shown in Fig. 4.1 (B). The estimated values in Figures 4.6 (B) and (C) suggest that the values of α and β may also be functions of x_1 . Based on the values in Figures 4.6 (B) and (C), we use the following two functions to approximate α and β , namely $\alpha = 3.25 + 7.5/x_1$ and $\beta = 11.8 + 8.2x_1$. Thus the final expression of the approximated C_2 is

$$C_2 = \frac{(3.25 + 7.5/x_1)(x_1(n-1) - y)}{11.8 + 8.2x_1 + y} - \frac{1}{y}. \quad (4.3.2)$$

To validate the proposed approach (4.3.2), we compare the optimal value of C_2 in Fig. 4.6 (A) with that determined by (4.3.2). Fig. 4.6 (D) shows the difference between these two values under different values of x_1 and y , which suggests that approach (4.3.2) provides accurate approximation to the optimal value of C_2 . In summary, the time delay for producing a product P via the process of multi-step reactions (4.2.1) is determined by (4.2.5, 4.2.6, 4.3.2).

4.3.3 Time delay of mRNA degradation

Next we apply our established state-dependent delay model to study the mRNA degradation process of gene ribosomal protein L30 (*RPL30*). Experimental studies have demonstrated the transcript decay dynamics of two constructs for this gene,

namely construct A-ACT1 UAS (upstream activating sequence) and construct B-RPL30 UAS (Bregman *et al.*, 2011). In experiments, mRNA molecule decay dynamics was monitored after blocking transcription by using drug 1, 10–phenanthroline (Bregman *et al.*, 2011). Therefore, it is assumed that no further transcription occurs after drug application. Since there is no explicit information regarding the mRNA copy number in experiments, we tested the case with initial total mRNA number $s_0 (= 100)$.

mRNA degradation has been modelled as a multi-component model that contains mRNA transcript synthesis, mRNA translocation, poly (A)-shortening process, and terminal deadenylation (Cao and Parker, 2003). A simplified model of multi-step reactions was proposed to put a number of terminal deadenylation reactions into a single reaction (Tian, 2014). Here we use the delayed reactions (4.2.3, 4.2.4) to represent the degradation dynamics, where X_1 is mRNA molecule with full length of poly(A)-tail and imaginary species G represents any one of the transcripts in the poly(A)-shortening process. The initial number of imaginary species y_0 and degradation rate k are unknown parameters that need to be estimated to match experimental data. In addition, the manifesting time of these imaginary species are uniformly distributed in time interval $[0, MT]$ and

$$MT = \text{delay}(x_{10}, y_0, k, n)/D, \quad (4.3.3)$$

where $\text{delay}(x_{10}, y_0, k, n)$ is the time delay determined by the initial system state (x_{10}, y_0) , degradation rate k , and number of steps n using the proposed method (4.2.5 ~ 4.3.2). We use the rejection method to search for the optimal parameters of y_0, k and D . Using an Approximate Bayesian Computation (ABC) rejection sampling algorithm (Turner and Zandt, 2012), we select 150 sets of model parameters and use the set with minimal error as our final estimation. Based on 1000 simulations, Fig. 4.2 shows that the state-dependent delay model is able

to provide accurate description of mRNA degradation dynamics for the two constructs of gene *RPL30*. Distributions of inferred parameters in Fig. 4.3 (A) and (D) suggest that $\sim 25\%$ of initial mRNA molecules are imaginary species, namely the transcripts in the poly(A)-shortening process. In addition, distributions of value D in Fig. 4.3 (B) and (F) suggest that the degradation time points of these shortened transcripts are distributed in an interval that is only $40 \sim 50\%$ of the normal time delay interval. Thus these imaginary species may already exist in the middle of the shortening process.

Simulation results in Fig. 4.2 are based on the assumption that $s_0 = 100$. The next question is whether the assumed initial total mRNA number influences estimated model parameters. To answer this question, we simulated the delay model using the same parameter (k, D) but rescaled y_0 and experimental data based on initial total mRNA $s_0 [= 10, 50, 150, 200]$. Simulation results in Fig. 4.7 show that our estimated parameters can also derive accurate simulations for various initial mRNA numbers.

4.3.4 Time delay in gene expression

We have successfully used our proposed method to simplify a multi-step reaction system. The next question is whether our method can be applied to more complex systems. To answer this question, we now study the dynamics of a cell cycle-regulated gene (e.g. *SWI5*) based on the measured changes in their mRNA turnover during the cell cycle (Trcek *et al.*, 2011). *SWI5p* is a transcription regulator of late mitosis genes and it was measured to degrade with a single half-life of 8 min (Wang *et al.*, 2002). In addition, *NDD1* (Nuclear Division Defective) is an essential gene for the expression of *SWI5*. It has been shown that over expression of *NDD1* enhances the expression of *SWI5* (Loy *et al.*, 1999). Its expression is tightly regulated during the cell cycle. The expression of gene *NDD1* peaks during the

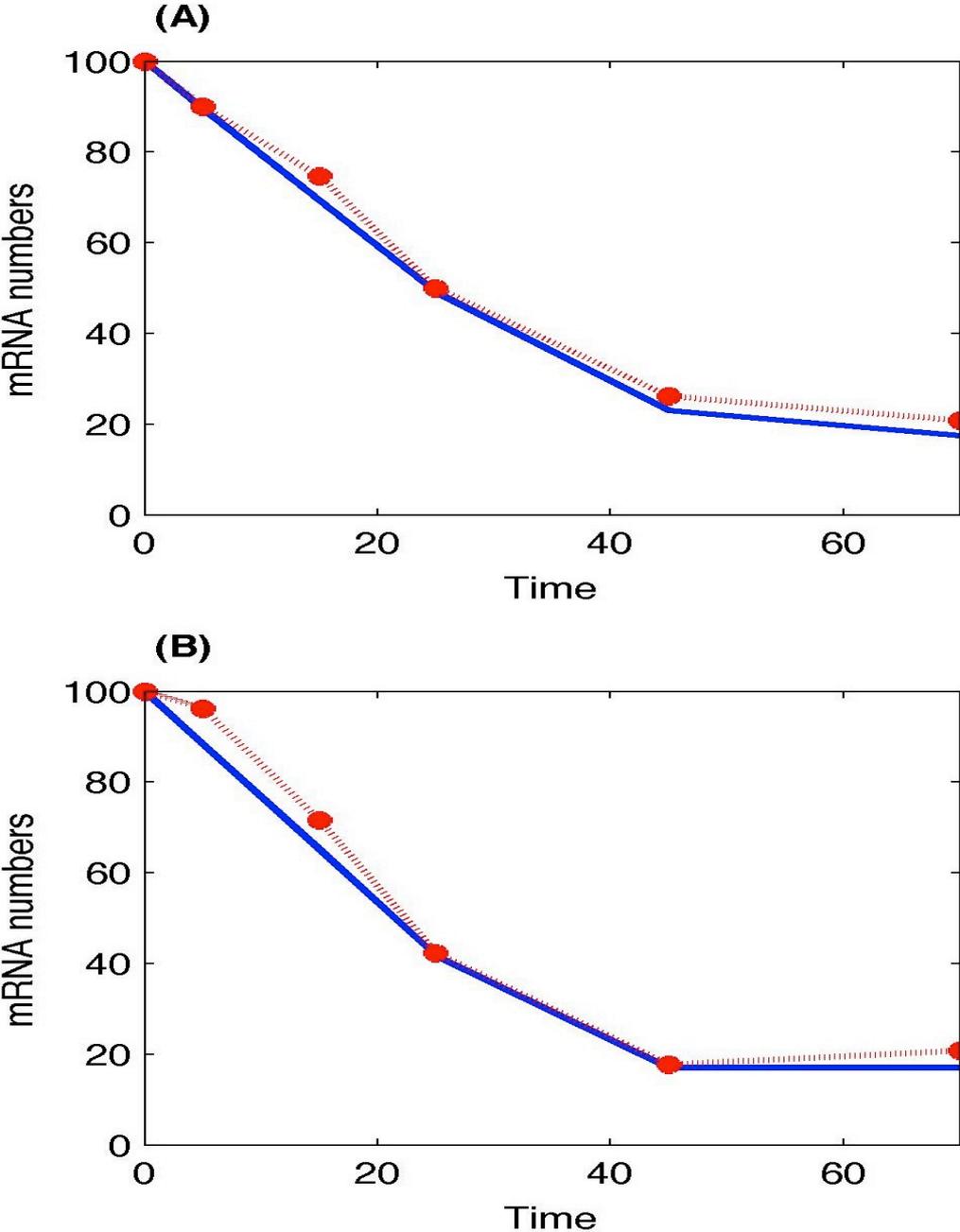


Figure 4.2: Simulation of mRNA degradation for gene RPL30 using the state-dependent delay model: (A) Construct ACT1 using estimated parameters $k = 0.1260, \gamma_0 = 23, D = 1.7184$. (B) Construct RPL30 using estimated parameters $k = 0.1260, \gamma_0 = 17, D = 1.7525$. (Solid line: averaged mRNA numbers based on 1000 simulations; dash-dot line: experimental data assuming $s_0 = 100$).

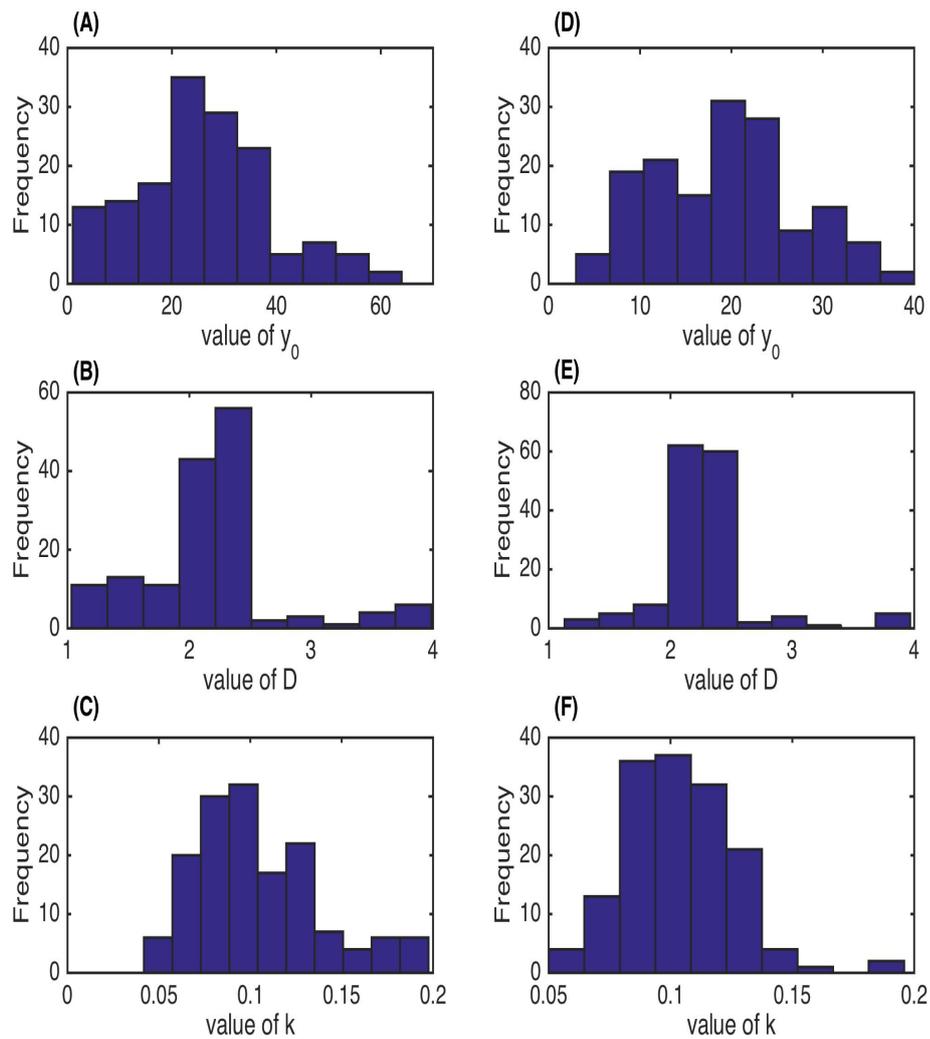


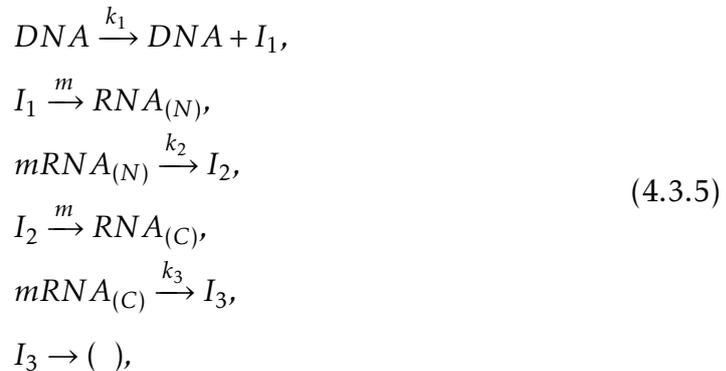
Figure 4.3: Distributions of estimated model parameters for gene RPL30 degradation: (A, B, C) Construct ACT1. (D, E, F) Construct RPL30.

S phase and is essential for expression of its target gene SWI5 during the G2/M phase (Loy *et al.*, 1999; Veis *et al.*, 2007).

A simple math model has been proposed to describe the expression of gene SWI5 based on experimental data measured in single cells. The degradation of mRNA molecules was described by a one-step reaction and simulation was used to measure the half-life of mRNA molecules (Trcek *et al.*, 2011). To accurately measure the half-life of mRNA transcripts, we propose a delayed model to describe the expression of genes SWI5. It is assumed that the transcription of this gene is activated by TF *NDD1*, which is realized by the rate of transcription

$$k_1 = \frac{a * [NDD1]}{b + [NDD1]}, \quad (4.3.4)$$

where a and b are parameters for genetic regulation. In addition, the elongation process needs time for RNAP II polymerase when travelling long the template DNA. Since not discussing the transcription process in details, we use a delay reaction with constant time delay for the synthesis of mRNA transcripts. Then mRNA transcripts translocate from nucleus to cytosol, and this process is also modelled by a delay reaction with constant time delay for simplicity. Finally mRNA molecules decay in cytosol via a multi-step process that is simplified as a state-dependent delay reaction, which is modelled using the proposed model (4.2.3,4.2.4). Thus the proposed model for the expression of gene SWI5 is given below



where $mRNA_{(N)}$, $mRNA_{(C)}$ are mRNA molecules in nucleus and cytosol; I_1, I_2, I_3 are imaginary species for $mRNA_{(N)}$, $mRNA_{(C)}$ and shortening mRNA, respectively. We use the inferred concentration of $[NDD1]$ in (Chen *et al.*, 2009), which is consistent with the drafted TF activity in (Trcek *et al.*, 2011), as the activity of this TF. In addition, experimental studies show that gene expression is regulated by mechanisms of cell cycle. In yeast, the mitosis process at ~ 49 min of each cell cycle will terminate the process of transcription. This regulatory mechanism was realized by the assumption that the activity of $[NDD1]$ is zero after 49 min of each cell cycle (Gandhi *et al.*, 2011).

The measured mRNA copy numbers in single cells are used to infer regulation parameters a, b , rate constant k_3 , and transcription and translocation delays. We use the ABC rejection sampling algorithm to search for optimal model parameters. Using simulation error to both cytosol and nucleus data as the criterion, we select 100 set of model parameters with small simulation error. The parameter set with the minimal error is the final inference result. Fig. 4.4 show that numerical simulations can match experimental data very well. In addition, the distribution of transcriptional time delay in Fig. 4.5 (A) are consistent with the experimental estimations showing that the time delay in transcription is ~ 35 min. An interesting observation is the degradation rate of mRNA is $\sim 0.1/\text{min}$, which suggests that the half-life of mRNA molecules is about 6.93 min.

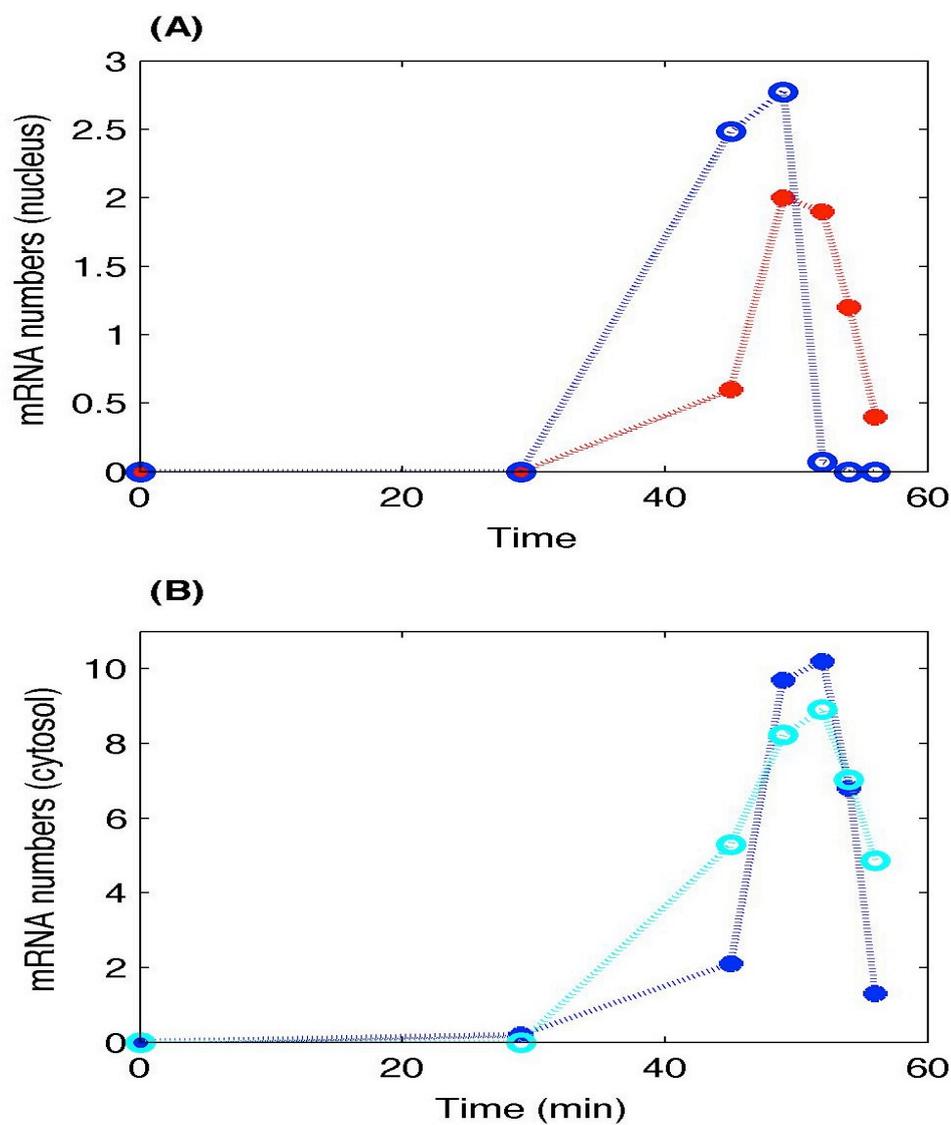


Figure 4.4: Simulation of gene transcription for gene SWI5 using the state-dependent delay model: (A) mRNA copy number in nucleus. (B) mRNA copy number in cytosol. (dot: experimental data; circle: simulations). Estimated parameters are $a = 3.9137$, $b = 8.2969$, $\tau_1 = 36.9431$, $\tau_2 = 2.0682$, $k_2 = 2247.9$, $k_3 = 1.0926$.

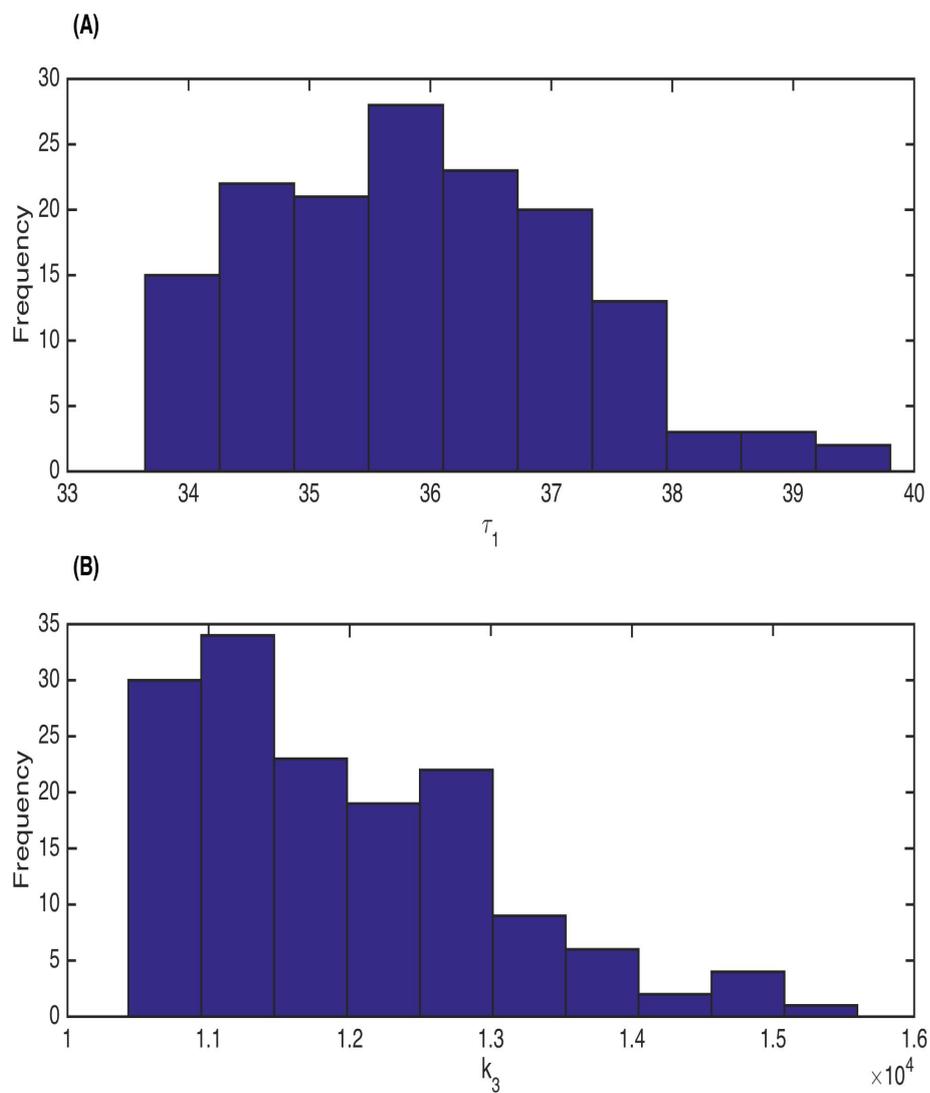


Figure 4.5: Distributions of estimated model parameters for gene SWI5 transcription: (A) Transcription delay. (B) Degradation rate constant k_3 .

4.4 Discussion and Conclusion

In this work, we propose a new algorithm to determine time delay in chemical reaction systems according to the system state. Using the process of multi-step reactions as the test problem, we utilize both the analytical solution of ODE model and stochastic simulation to determine the relationship between the system state and value of time delay. The proposed method is applied to model the degradation process of mRNA molecules based on experimental data measured in single cells. For the first test system of mRNA degradation, our model gives simulations with better accuracy comparing with existing modelling methods. For the second test system of gene expression, our model provides simulated dynamics with very good accuracy for both synthesis and degradation of mRNA transcripts. Simulation results in this work suggest that the proposed method is an effective approach to approximate multi-step reactions system more accurately.

Half-life is an important concept to measure the degradation process in biological studies. It is the amount of time required for a species from full amount to a half of the full amount as measured at the beginning of time period, based on the assumption that the quantity follows an exponential decay. However, for many biological molecules, the decay process may not be exponential; rather it follows multi-step reactions. Thus, molecules at the intermediate states are also important for determining the value of half-life. That may be the reason to explain the difference between the determined half-time under different experimental conditions. Using the inferred degradation rate in the state-dependent delay model, our results suggest that our calculated half-time of mRNA molecules are between the determined values in the published papers.

The advances in systems biology have raised more challenges for modeling large-scale molecular regulatory networks. Although a trend of mathematical modeling

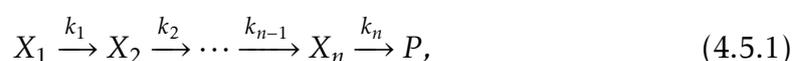
is to construct more and more mechanistically detailed models, the complexity of biological networks, lack of experimental data and requirement of computing power have put a limitation on the complexity of mathematical models. Recently various methods have been developed to reduce model complexity (Rao *et al.*, 2014; Mackey *et al.*, 2014). Simultaneously research works have been conducted to explore the conditions and assumptions of these simplified models in order to obtain accurate simulations (Thomas *et al.*, 2012; Schnell, 2014). This work represents a step in developing accurate delayed models for chemical reaction systems. More research work is strongly needed to study other types of multi-step reactions systems as well as the complex systems that include multi-step reactions processes as subsystems. These interesting problems will be potential topics of future research.

4.5 Supplementary Information

This supplementary information first gives a description of the exact solution of the multi-step chemical reactions in part (4.5.1). Then part (4.5.2) provides two algorithms to calculate the value of time delay using the stochastic simulation algorithm. Part (4.5.3) derives the formula to calculate time delay based on the analytical solution of the ordinary differential equation (ODE) model for the multi-step reaction system. Finally we give two supplementary figures for simulation results.

4.5.1 Multi-step chemical reaction system

The starting point of this study is the following system with a series of chemical reactions:



where X_i represents the i -th state of a molecule and k_i is the i -th reaction rate constant. Here \hat{A} is the product, which may also be “()” if it is a degradation process. Denote s as the total copy number of molecules in all states, given by

$$s = \sum_{i=1}^n x_i, \quad (4.5.2)$$

where x_i is the copy number of X_i state, and y as the sum of molecule numbers except x_1 , namely $y = \sum_{i=2}^n x_i$.

The dynamics of system (4.2.1) can be described using an ODE model as follows:

$$\begin{aligned} \frac{dx_1}{dt} &= -k_1 x_1, \\ \frac{dx_2}{dt} &= k_1 x_1 - k_2 x_2, \\ &\vdots \\ \frac{dx_n}{dt} &= k_{n-1} x_{n-1} - k_n x_n. \end{aligned} \quad (4.5.3)$$

For simplicity, it is assumed that $k_1 = k_2 = \dots = k_n = k$. The exact solutions of system (7.3.1) can be derived as

$$\begin{aligned} x_1 &= x_{10} e^{-kt} \\ x_2 &= x_{10} k t e^{-kt} + x_{20} e^{-kt} \\ x_3 &= \frac{x_{10}}{2} k^2 t^2 e^{-kt} + x_{20} k t e^{-kt} + x_{30} e^{-kt} \\ &\vdots \\ x_n &= \left[\frac{x_{10}}{(n-1)!} (kt)^{n-1} + \frac{x_{20}}{(n-2)!} (kt)^{n-2} + \dots + x_{n0} \right] e^{-kt}, \end{aligned} \quad (4.5.4)$$

where x_{i0} represents the initial copy number of X_i molecule at $t = 0$. Therefore, the total molecule number is represented by

$$s = s_0 e^{-kt} + (s_0 - x_{n0}) k t e^{-kt} + (s_0 - x_{(n-1)0} - x_{n0}) \frac{(kt)^2}{2!} e^{-kt} + \dots + \frac{x_{10}}{(n-1)!} (kt)^{n-1} e^{-kt}, \quad (4.5.5)$$

where $s_0 = (x_{10} + \dots + x_{n0})$. We assume that the initial conditions are x_{10} and $x_{20} = x_{30} = \dots = x_{n0} = \frac{y_0}{n-1}$. Then the total molecule number is represented by

$$s = e^{-kt} \left\{ (x_{10} + y_0) \left[1 + kt + \dots + \frac{(kt)^{n-1}}{(n-1)!} \right] - \frac{y_0 kt}{n-1} \left[1 + kt + \dots + \frac{(kt)^{n-2}}{(n-2)!} \right] \right\}, \quad (4.5.6)$$

which can be approximated by

$$s = e^{-kt} \left\{ (x_{10} + y_0) \left[e^{kt} - \frac{(kt)^n}{n!} e^{k\xi_1} \right] - \frac{y_0 kt}{n-1} \left[e^{kt} - \frac{(kt)^{n-1}}{(n-1)!} e^{k\xi_2} \right] \right\}, \quad (4.5.7)$$

When the number of reaction n is not small, such as the model of mRNA degradation with $n = 9$ (Wu *et al.*, 2013b), we can further assume that $\xi_1 = \xi_2 = \xi$.

4.5.2 Algorithm for calculating time delay

Stochastic simulation algorithm (SSA) is used to determine time delay based on various system conditions. Algorithm 1 is used to determine the time delay shown in Figure 4.1 (A). It calculates two waiting times and then the delay for each molecule.

Algorithm 1

- 1) The initial condition is $x_{10} > 0$ and $x_{i0} = 0$ for $i > 1$ at $t = 0$. The total initial total copy number is $s_0 = x_{10}$.
- 2) Calculate the value of propensity function $a_i = k_i x_i$ and $a_0 = \sum_{i=1}^n a_i$.

- 3) The waiting time of the next reaction is determined by

$$\mu = \frac{1}{a_0} \ln \frac{1}{r_1}, \quad (4.5.8)$$

where $r_1 \sim U(0, 1)$.

- 4) Generate a sample $r_2 \sim U(0, 1)$ to determine which reaction with index j from those multi-step reactions will occur.
- 5) Update the system by the determined reaction index j in step 4)

$$X(t + \mu) = X(t) + v_j, \quad (4.5.9)$$

if the reaction is the i -th first reaction $X_1 \rightarrow X_2$, then $\tau_{1i} = t + \mu$. If the reaction is the i -th last reaction $X_n \rightarrow P$, then $\tau_{2i} = t + \mu$.

- 6) Go to step 2.
- 7) Calculate time delay of the i -th molecules as $\tau_i = \tau_{2i} - \tau_{1i}$.

The following Algorithm 2 is used to calculate the delay of the first delay reaction based on different system states in Figure 4.1 (B).

Algorithm 2

- 1) The initial condition is $x_{10} > 0$ and $y_0 > 0$. First determine the values of (x_{20}, \dots, x_{n0}) based on the value of y_0 .
- 2) Calculate values of propensity function $a_i = k_i x_i$ and $a_0 = \sum_{i=1}^n a_i$.
- 3) The waiting time of the next reaction is determined by

$$\mu = \frac{1}{a_0} \ln \frac{1}{r_1}, \quad (4.5.10)$$

where $r_1 \sim U(0, 1)$.

- 4) Generate a sample $r_2 \sim U(0, 1)$ to determine which reaction with index j from those multi-step reactions will occur.
- 5) Update the system by determined reaction in step 4)

$$X(t + \mu) = X(t) + v_j, \quad (4.5.11)$$

if the reaction is the first of first reaction $X_1 \rightarrow X_2$, then $\tau_1 = t + \mu$. If the reaction is the first of last reaction $X_n \rightarrow P$, then $\tau_2 = t + \mu$.

- 6) Stop the algorithm after the first of the last reaction, and calculate the time delay $\tau = \tau_2 - \tau_1$.

4.5.3 Formulation of time delay

We use a reaction with time delay to simplify system (4.2.1), which is described as follows:



Here reaction (4.5.12) is the first reaction of system (4.2.1), while delayed reaction (4.5.13) is a simplification of the process from state X_2 to product P . The time delay in reaction (4.5.13) is the sum of waiting time experiencing $n - 1$ consecutive reactions from state X_2 to product P . Thus the imaginary state G represents any one of the intermediate states X_2, \dots, X_n and its molecular number is $y = \sum_{i=2}^n x_i$.

We need to determine the time delay based on current system state (X_1, G) , When the first reaction fires, a molecule from X_1 moves into the queue structure of time delay L in which there are already y imaginary molecules. When the newly added

molecule turns to product P , it is assumed that all y molecules queued before the newly added molecule already turn into the product. Considering the time for the first molecule from X_1 -state molecules to turn to product, the total molecule number should be reduced from $x_1 + y$ to $x_1 - 1$. The time delay is defined as

$$\tau = \tau_2 - \tau_1, \quad (4.5.14)$$

where τ_1 is the firing time of the first reaction $X_1 \rightarrow X_2$, and τ_2 is the firing time of the last reaction $X_n \rightarrow P$ and the system state after update is $s = x_1 - 1$.

We use computational simulations to determine the value of time delay. The value τ_1 is determined by stochastic simulation algorithm. The key issue is to determine the value of τ_2 . Given the system state as (x_1, y) at time t , the time t for the first X_1 molecule turns into product P is

$$x_1 - 1 = (x_1 + y) - \frac{ykt}{n-1} + \frac{y(kt)^n}{n!} e^{k(\xi-t)} \left[\frac{ny}{n-1} - (x_1 + y) \right], \quad (4.5.15)$$

which can be simplified as

$$e^{k(\xi-t)} (kt)^n \left(x_1 + y - \frac{ny}{n-1} \right) = \left[1 + y \left(1 - \frac{kt}{n-1} \right) \right] n!. \quad (4.5.16)$$

Denote

$$\begin{aligned} C_1 &= x_1 + y - \frac{ny}{n-1}, C_2 = 1 - \frac{kt}{n-1} \\ C &= \frac{1 + C_2 y}{C_1} n!. \end{aligned} \quad (4.5.17)$$

Equation (4.5.16) is simplified as

$$e^{k(\xi-t)} (kt)^n = C. \quad (4.5.18)$$

There are a number of undetermined coefficients in (4.5.18). Thus, we first use a special case to determine the value of ξ by letting $y = 0$. Thus the coefficients in the exact solution are x_1 and $x_2 = \dots = x_n = 0$, and the total molecule number is

$$s = x_1 e^{-kt} \left[1 + kt + \dots + \frac{(kt)^{n-1}}{(n-1)!} \right], \quad (4.5.19)$$

which can be approximated by ($0 < \xi < t$)

$$s = x_1 e^{-kt} \left[1 - e^{k(\xi-t)} \frac{(kt)^n}{n!} \right]. \quad (4.5.20)$$

To find the time τ_2 taken for the total copy number s to change from x_1 to $x_1 - 1$, we rewrite the equation (4.5.16) as

$$e^{k(\xi-t)} t^n = \frac{n!}{x_1 k^n}. \quad (4.5.21)$$

The solution of the time τ_2 in the above equation can be represented by

$$\tau_2 = -\frac{n}{k} W \left[-\frac{1}{n} \left(\frac{e^{-k\xi} n!}{x_1} \right)^{\frac{1}{n}} \right], \quad (4.5.22)$$

where $W(x)$ is the Lambert W function. To determine an optimal value of $\xi (\in (0, t))$, we compared the time delays $\tau_2 - \tau_1$ obtained using (4.5.7) with three values ($\xi = (0, 0.5, 1)t$) with those obtained from stochastic simulations using SSA. We found that, when $\xi = \frac{t}{2}$, the formula (4.5.22) provides more accurate estimate for time delay. Thus we use the following formula, given by

$$\tau_2 = -\frac{2n}{k} W \left[-\frac{1}{2n} \left(\frac{n!}{x_1} \right)^{\frac{1}{n}} \right]. \quad (4.5.23)$$

Now we return to the general case when $\gamma > 0$. First we consider a particular case with $C_1 = 0$. Then the left hand side of equation (4.5.16) is zero. Then we have

$$\tau_2 = -\frac{1 + x_1 n - x_1}{k x_1}. \quad (4.5.24)$$

Otherwise, the right-hand side of equation (4.5.18) is always positive. The solution of equation (4.5.18) in terms of t is represented by a Lambert W function. Using the optimal value $\xi = \frac{t}{2}$, the time to reach the system state with $x_1 - 1$ molecules is

$$\tau_2 = -\frac{2n}{k} W\left[-\frac{1}{2n} C^{\frac{1}{n}}\right]. \quad (4.5.25)$$

In summary, we have an expression for the time delay

$$\tau = \tau_2 - \tau_1, \quad (4.5.26)$$

where the value τ_1 is determined by stochastic simulation algorithm, and

$$\tau_2 = \begin{cases} \frac{1+x_1 n-x_1}{k x_1} & \text{if } C_1 = 0 \\ -\frac{2n}{k} W\left(-\frac{1}{2n} C^{\frac{1}{n}}\right) & \text{if } C_1 \neq 0, \end{cases} \quad (4.5.27)$$

Note that the value of C in solution (4.5.27) actually depends on future time point t , which is unknown. A formula is needed to approximate the value of C by numerical simulation, which will be studied in the paper.

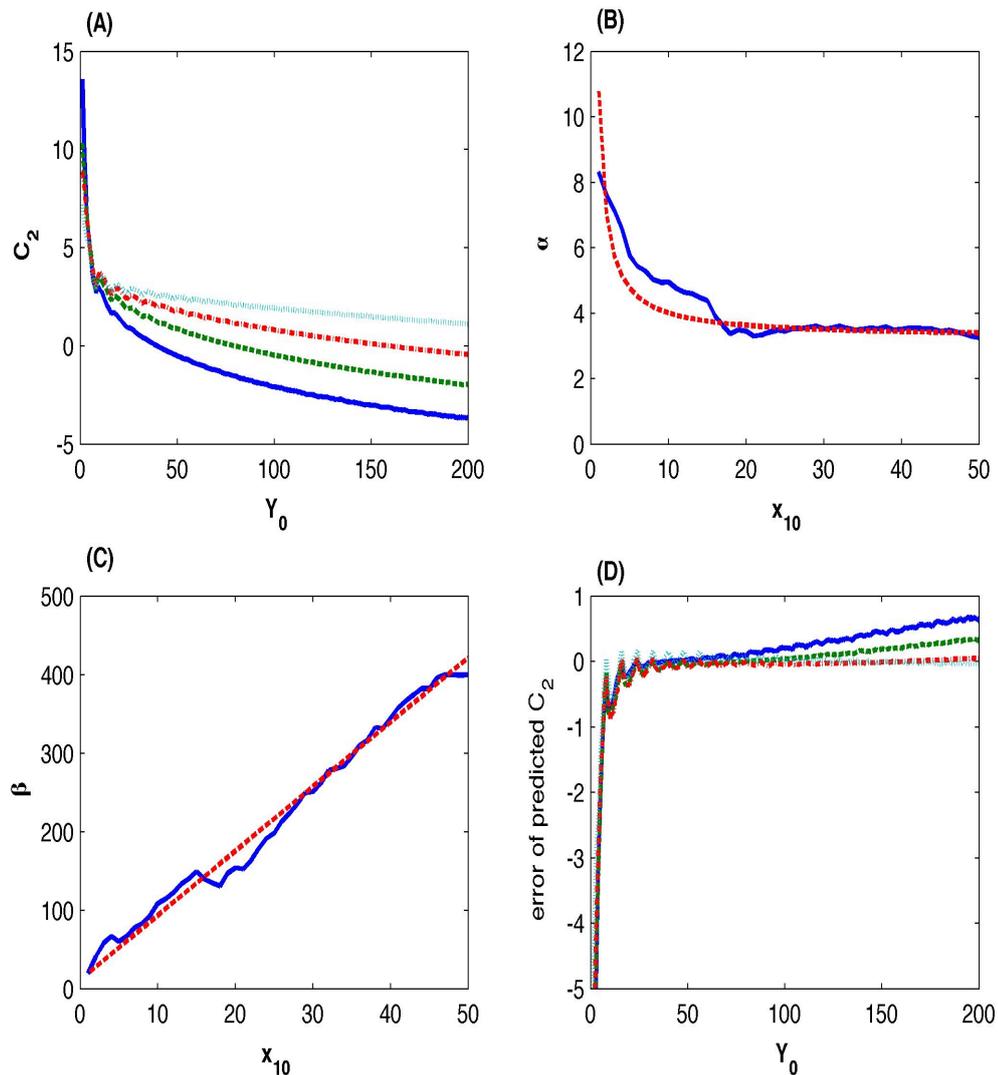


Figure 4.6: A new algorithm for calculating time delay that is dependent on system state: (A) Estimated optimal values of C_2 based on different system states (x_{10}, y_0) that match time delay showing in Figure 4.1 (B). Each line represents the optimal value of C_2 for a particular value of x_{10} . For a fixed value of y_0 , the smaller the value of x_{10} is, the smaller the value of C_2 becomes. (Solid-line: $x_{10} = 5$, dash-line: $x_{10} = 10$, dash-dot line: $x_{10} = 20$, dot-line: $x_{10} = 50$). (B) Values of α . (blue-solid line: estimated values based on simulated time delay in Figure 4.1 (B); red-dash line: prediction from $\alpha = 3.25 + 7.5/x_1$). (C) value of β . (blue-solid line: estimated values based on simulated time delay in Figure 4.1 (B); red-dash line: prediction from $\beta = 11.8 + 8.2x_1$). (D) The difference between the predicted values of C_2 and optimal values of C_2 in Figure 4.6 (A). (Solid-line: $x_{10} = 5$, dash-line: $x_{10} = 10$, dash-dot line: $x_{10} = 20$, dot-line: $x_{10} = 50$).

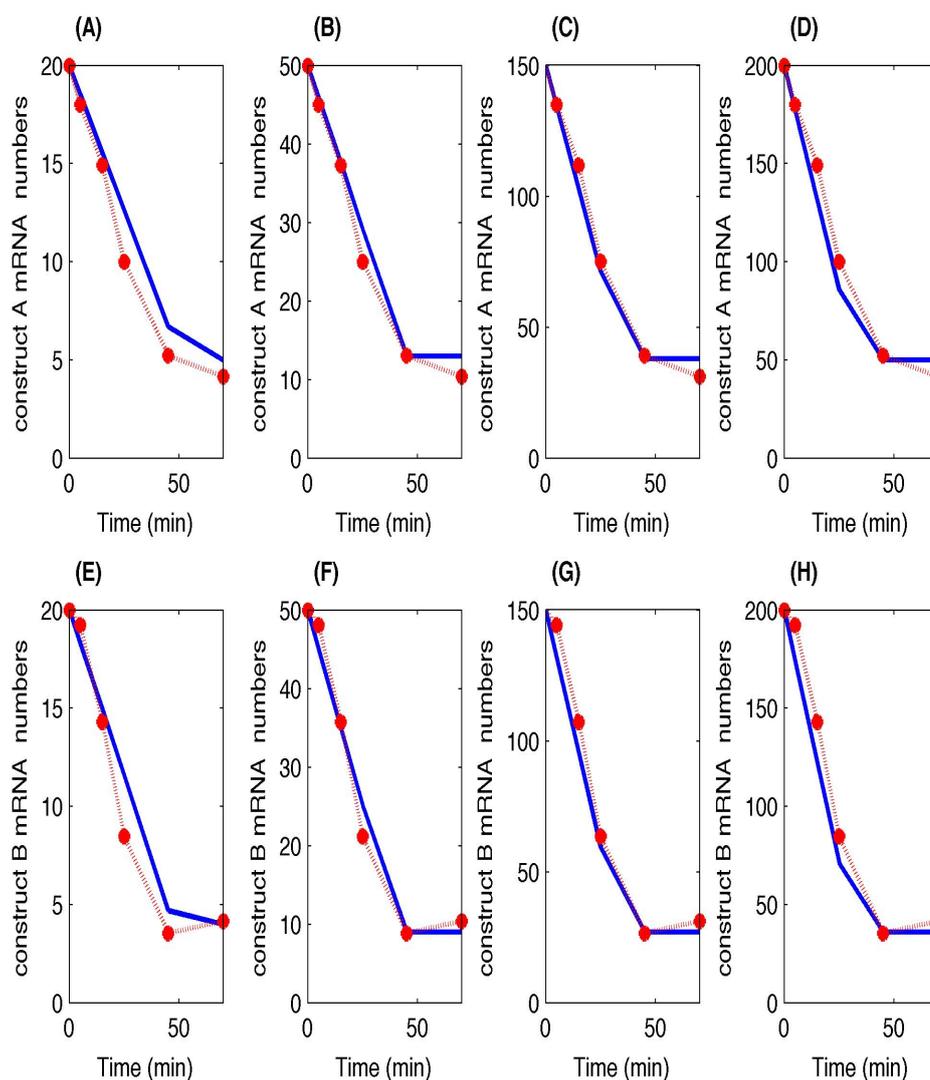


Figure 4.7: Simulation of mRNA degradation for gene *RPL30* using different initial mRNA number s_0 . For each construct, parameters k and D are the same as those in Figure 4.2. The value of y_0 is proportional to the value of s_0 : (A, B, C, D) Construct ACT1. (E, F, G, H) Construct RPL30. (A, E) $s_0 = 20$. (B, F) $s_0 = 50$. (C, G) $s_0 = 150$. (D, H) $s_0 = 200$. (Solid-line: simulation. Dash-dot line: experimental data).

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 5

Declaration by candidate

In the case of Chapter 5, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Developed, established and verified the method Wrote programming codes and the article	90%

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Kate Smith-Miles	Provided helpful guidance and proofreading	
Tianhai Tian	Provided helpful guidance and editorial work	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature		Date 29/04/15
Main Supervisor's Signature		Date 29/04/15

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 5

Approximate Bayesian Computation for Estimating Rate Constants in Biochemical Reaction Systems

Chapter 5 is based on the article Wu Q, Smith-Miles K, Tian T. 2013a. Approximate bayesian computation for estimating rate constants in biochemical reaction systems. In: Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on. pp. 416–421, doi: 10.1109/BIBM.2013.6732528.

Abstract. *To study the dynamic properties of complex biological systems, mathematical modelling has been used widely in systems biology. Apart from the well-established knowledge for modelling techniques, there are still some difficulties while understanding the dynamics in system biology. One of the major challenges is how to infer unknown parameters in mathematical models based on the experimentally observed data sets. This is extremely difficult when the experimental data are sparse and the biological systems are stochastic. To tackle this problem, in this work we revised one computation method for inference called approximate Bayesian computation (ABC) and conducted extensive computing tests to examine the influence of a number of factors on the performance of ABC. Based on simulation results, we found that the number of stochastic simulations and step size of the observation data have substantial influence on the estimation accuracy. We applied the ABC method to two stochastic systems to test the efficiency and effectiveness of the ABC and obtained promising approximation for the unknown parameters in the systems. This work raised a number of important issues for designing effective inference methods for estimating rate constants in biochemical reaction systems.*

References are considered at the end of the thesis.

Chapter 5

Approximate Bayesian Computation for Estimating Rate Constants in Bio- chemical Reaction Systems

5.1 Introduction

Studying complex biological systems with quantitative methods has drawn more and more attentions in systems biology in recent years. Building mathematical models is one of the most widely used methods to investigate the dynamic properties of biological systems among various research approaches. In particular, it requires more knowledge and condensation of assumptions to construct mathematical models for complex biological systems for genetic regulatory networks and cell signalling pathways into simple coherent frameworks. Simple mathematical models can be applied to make testable predictions, which can be used for biologists to confirm the predictions and design new experiments. With the new data obtained from newly designed experiments, it can be further used to improve the established mathematical models (Ashyraliyev *et al.*, 2009).

While establishing the models, there are two major steps: one is to determine the basic structure that describes the system and the other is to estimate the unknown parameters in the model (Zhan and Yeung, 2011). Generally, we do not have adequate information to measure the unknown parameters for the vast majority of systems and especially for biological systems. Moreover, the given information especially from biological experimental data is often scarce and incomplete, and the likelihood surfaces of large models are complex. It has been treated as one of the key questions in systems biology to solve for the unknown parameters within any model structure and it is often referred to as a reverse engineering problem (Kikuchi *et al.*, 2003; Tsai and Wang, 2005). Therefore, to help analyze those biological dynamical systems, new and effective inference methods are required.

Among extensive research that has been conducted for the development of inference methods during the last decade, one of the major approaches is Bayesian inference methods. Bayesian inference is the method where the Baye's rule is applied to update a probability estimate. The main advantage of Bayesian inference is the ability for inferring the whole probability distributions of the parameters, rather than just a point estimate. It is able to extract useful information from noisy or uncertain data (Wilkinson, 2007), where this includes both measurement noise and intrinsic noise that is critical in biological systems with species of low copy numbers (McAdams and Arkin, 1999). Also, handling estimations for stochastic systems using Bayesian inference methods is more robust as for deterministic systems (Toni *et al.*, 2009). However, there still exist some disadvantages of this approach such as the computational time, while using analytical approaches are not feasible for non-trivial problems. Nonetheless, developments for overcoming various difficulties have taken place during the last twenty years and the techniques that have been established most recently in Bayesian inference methods include the Markov chain Monte Carlo (MCMC) techniques, ensemble methods, and sequential Monte Carlo (SMC) methods that do not require likelihoods. All

these techniques have been successfully applied to biological systems, but usually for the systems that involve lower-dimensional problems or with a relatively large number of experimental data (Battogtokh *et al.*, 2002; Sisson *et al.*, 2007).

When the likelihoods for parameter estimations are computationally intractable, besides the above Bayesian techniques, we can apply the recently developed algorithm called approximate Bayesian computation (ABC) frameworks (Beaumont *et al.*, 2002; Marjoram *et al.*, 2003). The ABC methods admit realistic inference on problems that were intractable only a few years ago. Extensive studies that lead to a substantial methodological advance suggested that the ABC methods yield reliable parameter estimates with credible intervals and is relatively computationally efficient. This method can be applied to various models, allows for discrimination among sets of candidate models in a formal Bayesian model selection sense, and gives us an assessment of parameter sensitivity. Unlike the usual Bayesian methods, the ABC methods evaluate the likelihood based on simulations through the comparison between the observed data and simulated data (Pritchard *et al.*, 1999). The rapidly increasing application of the ABC methods has been seen in a diverse range of fields, including molecular genetics, ecology, epidemiology, evolutionary biology, and extreme value theory (Marjoram and Tavaré, 2006; Butler *et al.*, 2006; Tanaka *et al.*, 2006; Thornton and Andolfatto, 2006).

Even though a large number of studies have been done to apply the ABC method to deterministic models, limited research work has been carried out so far for the inference of stochastic models. Different from deterministic models, there are many open problems in the inference of stochastic models, such as selection of objective function, influence of particle size and simulation number, choice of threshold values, etc.. Thus in this work, we undertook extensive computing experiments to examine the influence of a number of factors on the performance of the ABC methods. The remaining part of this paper is organized as follows. Section 5.2 describes the basic algorithm in detail as well as the ABC SMC algorithm.

Section 5.3 part 5.3.1 uses a simple chemical system to examine the influence of a number of factors on the performance of the ABC SMC algorithm. Section 5.3 part 5.3.2 uses an auto-regulatory gene network system to show the ability of the ABC methods to infer parameters in a larger system.

5.2 Method

The ABC method, which is a computational simulation technique, aims at inferring posterior distributions where likelihood functions are not easy to compute. Its high efficiency is a result of replacing the calculation of the likelihood function by a comparison between the observed and simulated data. For inference problems, we usually start with a set of experimental data \mathbf{X} and let θ be the parameter vector to be estimated. An initial guess called prior distribution $\pi(\theta)$ for θ is assumed and we want to approximate the posterior distribution $\pi(\theta|\mathbf{X})$ given the data \mathbf{X} .

Based on previous research (Toni *et al.*, 2009), all ABC algorithms have the following major steps.

- Sampling: sample a candidate parameter θ^* from the proposed prior distribution $\pi(\theta)$.
- Simulation: simulate the results \mathbf{Y} of the proposed model with parameter θ^* .
- Comparison: compare the simulated data \mathbf{Y} with the observed data \mathbf{X} and find the distance $d(\mathbf{X}, \mathbf{Y})$ between them.
- Decision making (Selection): for a given tolerance or threshold value ϵ , accept the sampled parameter θ^* if $d(\mathbf{X}, \mathbf{Y}) \leq \epsilon$, otherwise, reject it and return to the first step of sampling.

With sufficient number of iterations for the above algorithm, we are able to obtain a set of estimated parameters within satisfaction and then the posterior distribution can be estimated with the distribution $\pi(\theta|d(\mathbf{X}, \mathbf{Y}) \leq \epsilon)$. However, the difficulties are how to define a suitable distance function for calculating the difference and to choose an optimal tolerance value. If the tolerance ϵ is sufficiently small, our obtained distribution will be a good approximation, which may be too costly to evaluate. If ϵ is large, we would obtain a distribution which may be useless for approximation.

Based on the generic form of the ABC algorithm, many methods have been developed including the ABC rejection sampler, which is a similar derivation as the above algorithm, and the ABC MCMC. The ABC MCMC algorithm solves the problem for long computing time due to a badly chosen prior distribution that is far away from posterior distribution. However, as the ABC MCMC introduces a concept of acceptance probability during the decision making step, then candidate parameters must meet two criteria at the same time. This will result in getting stuck in the regions of low probability and we may never be able to get a good approximation.

To avoid the problem raised using the ABC MCMC algorithm, the idea of particle filtering has been introduced. Instead of having one parameter vector at a time, we sample from a pool of parameter sets simultaneously and treat each parameter vector as a particle. The algorithm starts from sampling a pool of N particles for parameter vector θ through prior distribution $\pi(\theta)$. The sampled particle candidates $(\theta_1^*, \dots, \theta_N^*)$ will be chosen randomly from the pool and we will assign each particle a corresponding weight w to be considered as the sampling probability. For the first iteration, we assume that it has a equal weight of $\frac{1}{N}$ for each sampled particle. A perturbation and filtering process will be followed through a transition kernel $q(\cdot|\theta^*)$ to form a new set of particles θ^{**} . Similarly, using θ^{**} , data \mathbf{Y} can

be simulated and compared with experimental data \mathbf{X} . Then we adapt the same concept as the decision making step to choose the ideal estimated parameters.

A number of algorithms have been developed using the particle filtering technique, such as the partial rejection control, population Monte-Carlo and SMC. Each of them differs in the formation of weight w and the transition kernels $q(\cdot|\theta^*)$ they choose. We will only present here the SMC sampling method in detail which can be applied for stochastic biological systems.

The ABC SMC algorithm is a special case of sequential importance sampling (SIS) algorithm (Toni *et al.*, 2009). The algorithm is described in detail as follows.

Algorithm

1. Given data $\mathbf{X} = \{X_0, X_1, \dots, X_n\}$ at time points $t = [t_0, t_1, \dots, t_n]$ and any assumed prior distribution $\pi(\theta)$, define a set of threshold values $\epsilon_1, \dots, \epsilon_K$.

2. For iteration $k = 1$,

(a) Set the particle indicator $i = 1$, sample $\theta^* \sim \pi(\theta)$.

(b) Generate data \mathbf{Y} B_k times using θ^* .

(c) For $m = 1, \dots, B_k$, calculate the value of discrepancy $d(\mathbf{X}, \mathbf{Y}_m)$ and test for

$$|\mathbf{X} - \mathbf{Y}_m| \leq \alpha \mathbf{X},$$

where α is usually a constant value of 0.05.

If it is true, let $b_m(\theta^*) = 0$, otherwise it is one.

(d) Calculate

$$\epsilon = \sum_{m=1}^{B_k} b_m(\theta^*).$$

If $\epsilon < \epsilon_k$, update $\theta_i^k = \theta^*$ and move to the next particle $i = i + 1$.

(e) Assign weight $w_i^k = \frac{1}{N}$ for each particle.

3. Determine the variance for the particles in the first iteration

$$\sigma_1 = \sqrt{\text{var}(\theta_{\{1:N\}}^1)}$$

4. For iteration $k = 2, \dots, K$

(a) Start with $i = 1$, Sample $\theta^* \sim \theta_{i:N}^{k-1}$ using the calculated weights $w_{i:N}^{k-1}$.

(b) Perturb θ^* through sampling $\theta^{**} \sim N(\theta^*, \sigma_{k-1}^2)$ or $\theta^{**} \sim U(a, b)$, where value of a, b depends on θ^* and σ_{k-1}^2 .

(c) Generate simulations and calculate the error ϵ using the same steps as in 2(b) ~ (d).

(d) Assign weights

$$w_i^k = \frac{\pi(\theta_i^k) b_k(\theta_k^i)}{\sum_{j=1}^N w_j^{k-1} q(\theta_j^{k-1} | \theta_j^k, \sigma_{k-1})}$$

for each particle.

(e) If $i = N$, determine the variance for the particles in the first iteration

$$\sigma_k = \sqrt{\text{var}(\theta_{\{1:N\}}^k)}$$

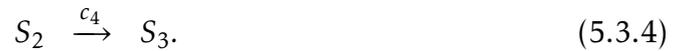
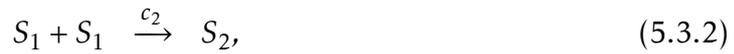
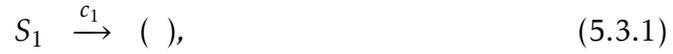
and update $k = k + 1$.

5.3 Results

5.3.1 Decay-dimerization model

The first system we tested is the model of reaction system that involves species decay and dimerization. This model begins with the first three reactions

(6.2.1,6.2.1,6.2.1), in which the dimerization step (6.2.1,6.2.1) is reversible (Daigle *et al.*, 2012). By adding a conversion reaction (6.2.1) to the reversible model, we have the system described as follows:



We start with an initial condition with $\mathbf{S} = (10000, 0, 0)$ and rate constants of $\mathbf{c} = (0.1, 0.002, 0.5, 0.04)$, which is termed as the exact rate constants in this test. The stochastic simulation algorithm (SSA) was used to simulate the stochastic system (Gillespie, 1977). A single trajectory data for this model during a period of $T = 30$ in a step size of $\Delta t = 1$ is presented in Fig. 5.1. This figure shows the dynamics of the system in which S_1 decreases sharply while S_2 increases in the beginning, starts to decrease steadily and S_3 increases gradually.

When applying the Algorithm described in previous section to estimate model parameters, we assumed the prior distribution for each estimated parameter follows a uniform distribution $\pi(\theta) \sim U(0, A)$ and for rate constants $c_1 \sim c_4$, the values of A are $(0.5, 0.005, 1, 0.1)$ (Tian *et al.*, 2007b).

Fig. 5.2 gives probability distributions of the estimated rate constant of c_4 over iterations ($2 \sim 5$). We choose a fixed value of 100 for particle number for all tests. In this test, the step size of the data is $\Delta t = 3$ and simulation number is $B_k = 100$. From this figure, it can be found that the probability distribution for c_4 starts from nearly a uniform distribution in the second iteration (Fig. 5.2A) and gradually converges to a normal distribution with a mean value that is close to the exact rate constant.

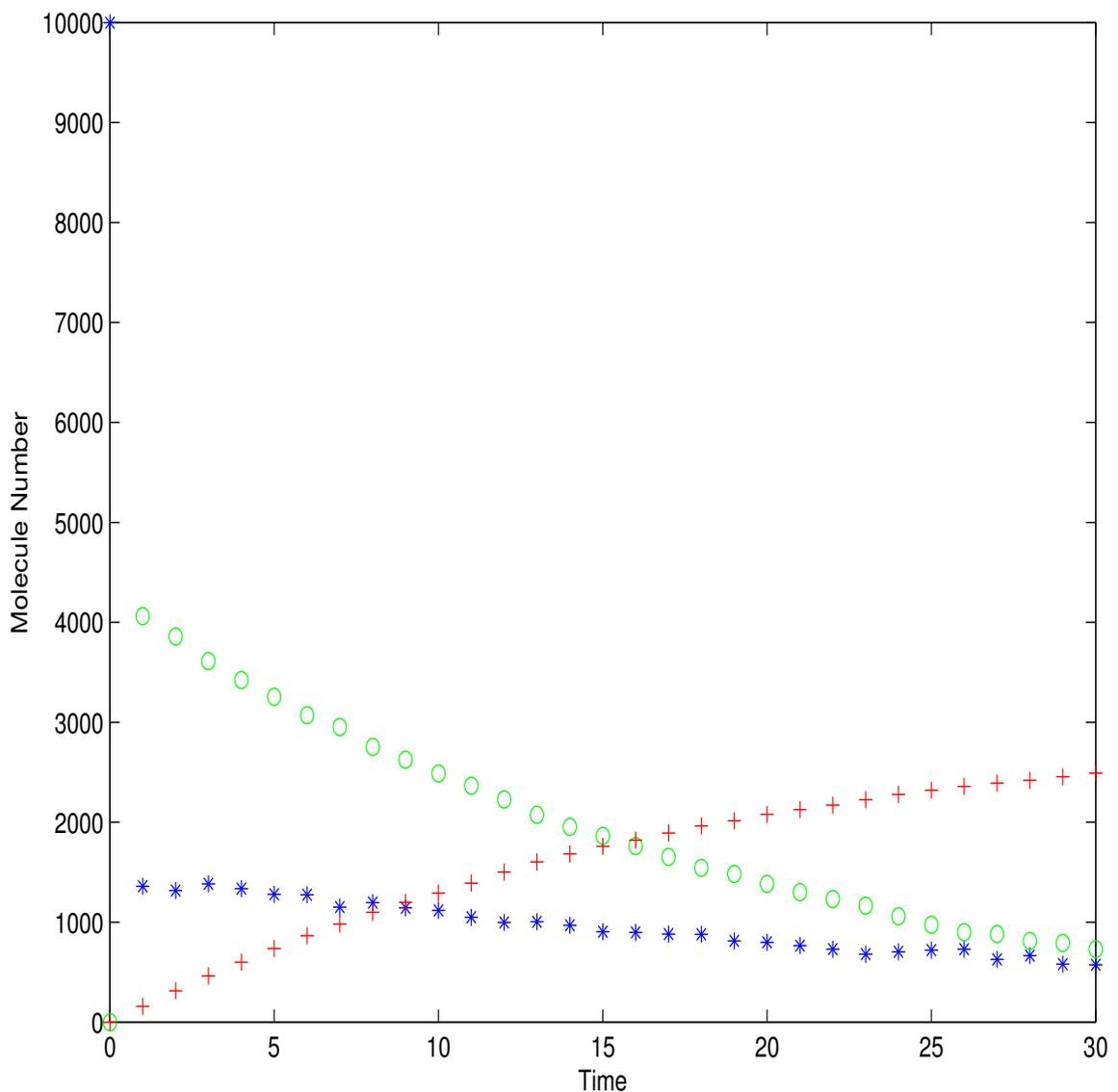


Figure 5.1: Simulated experimental data for system dynamics in a time length of 30 with step size Δt of 1 (Blue star for S_1 , green circle for S_2 , and red cross for S_3).

To examine the factor that reveals the convergence rate of particles over iterations, we define and calculate the mean count number for each iteration, which is the averaged count number of selections before accepting all one-hundred estimated parameter sets. We also define the averaged error by the sum of relative errors of each rate constant for each iteration. Fig.5.3 shows some examples for the values of mean count number and averaged error with a simulation number of 100 and step size of 3 and 5, respectively. We have also explored the averaged error

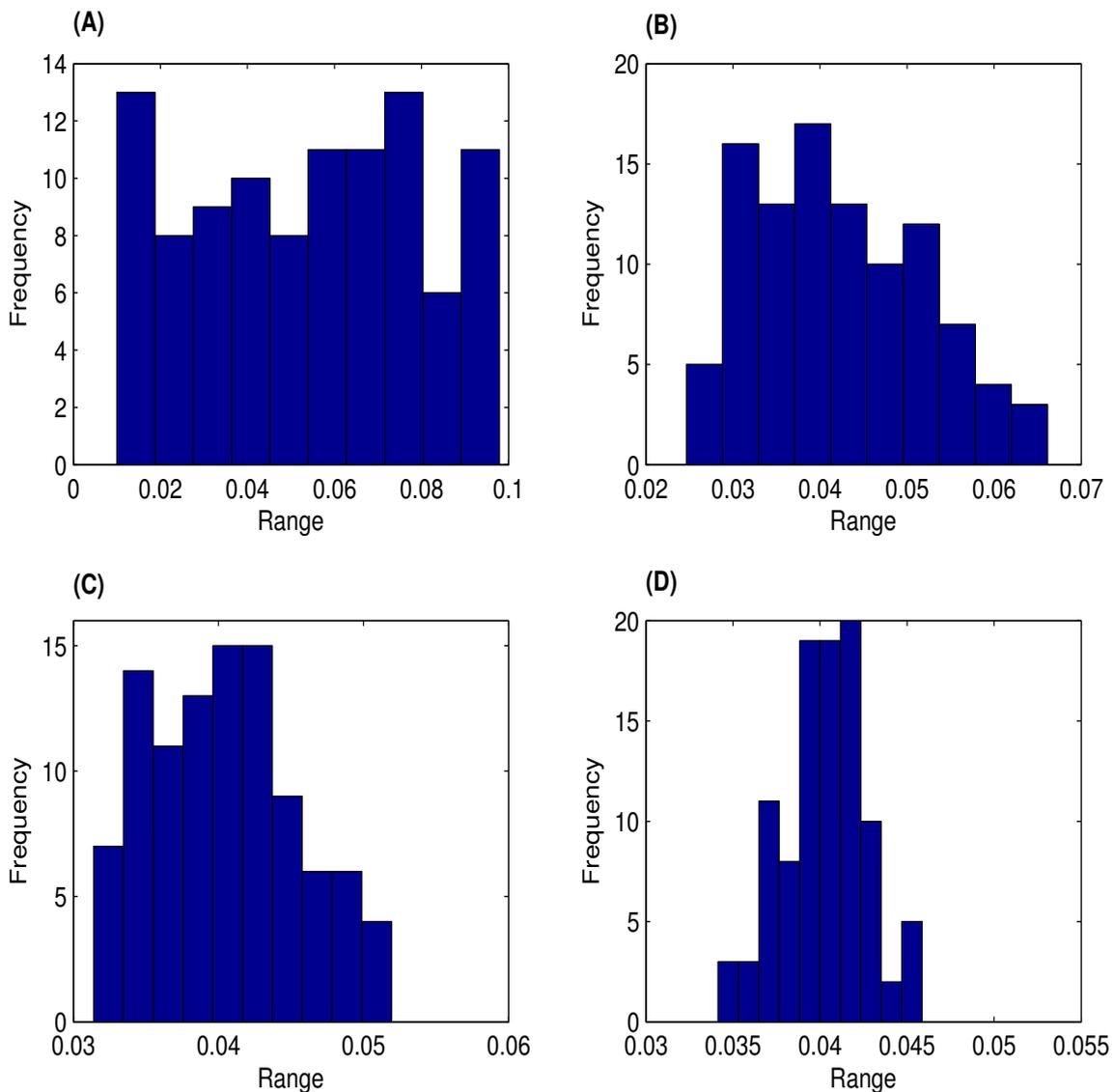


Figure 5.2: Probability distributions of estimated rate constant of c_4 over four iterations ((A): Iteration 2; (B): 3; (C): 4; (D): 5).

and mean count number under other test conditions with different simulation numbers and step sizes. Numerical results are consistent with those shown in Fig. 5.3, which suggests that mostly the averaged error decreases when the mean count number increases over five iterations.

Table 5.1 provides the averaged error of estimates that were obtained using different simulation numbers and step sizes. These numerical results indicate that when the simulation number is larger, we can obtain more stable estimates whose

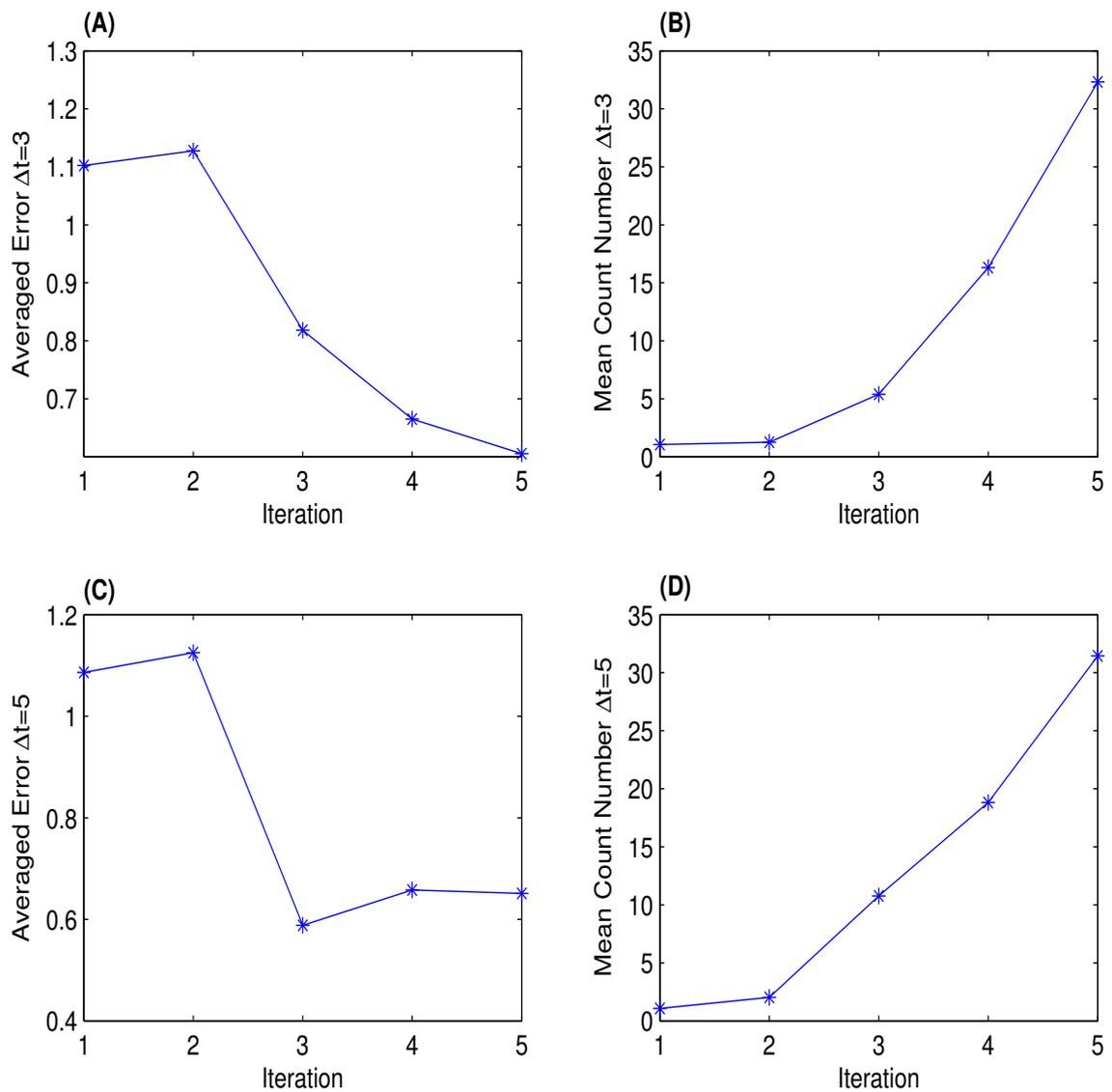


Figure 5.3: The averaged error of estimated parameters and mean count number of iterations with step size Δt of 3 ((A), (B)) and 5 ((C), (D)).

averaged errors have less fluctuations over iterations, which is consistent with the observations in (Tian *et al.*, 2007b). Certainly the cost of this stability is the large computing time required for stochastic simulation. An initial observation from Table 5.1 is that step size Δt has not much influence on the averaged error. However, after looking at the mean count number in Table 5.2, the estimates using larger step size were obtained at the cost of a much larger mean count number. Combining the observations in Fig. 5.3, we conclude that, if we use the same

computing time, the ABC algorithm will infer parameters with better accuracy when the step size is smaller, which is also consistent with the observations in (Tian *et al.*, 2007b).

Table 5.1: Comparison of averaged error for estimated rate constants over five iterations with different simulation numbers and step sizes

Simulation number	Data step size	Iteration number				
		1	2	3	4	5
100	1	1.050	0.995	1.177	0.663	0.610
	2	1.008	1.090	0.838	0.758	0.712
	3	1.102	1.127	0.818	0.665	0.605
	5	1.086	1.125	0.588	0.657	0.651
1000	1	1.019	0.925	1.007	0.864	0.745
	2	0.937	1.157	0.968	0.624	0.704
	3	1.012	1.132	0.851	0.678	0.741
	5	0.948	1.109	0.708	0.641	0.724
2000	1	1.167	0.990	1.093	0.752	0.680
	2	1.175	1.062	0.907	0.674	0.651
	3	1.118	1.224	0.900	0.661	0.672
	5	0.873	0.999	0.628	0.596	0.627

5.3.2 Prokaryotic auto-regulatory gene network

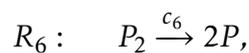
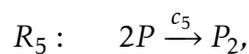
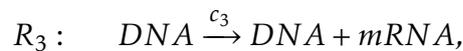
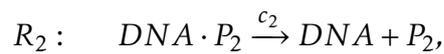
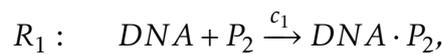
After successfully estimating the parameters in the system of four reactions, we continue the inference study for a more complex model. The second system we tested is a prokaryotic auto-regulatory gene network. In this reaction system, it involves both transcription and translation. In addition, dimers of the protein suppress its own gene transcription by binding to a regulatory region upstream of

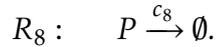
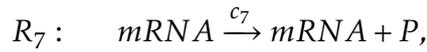
Table 5.2: Comparison of mean count number for estimated rate constants over five iterations with different simulation numbers and step sizes

Simulation number	Data step size	Iteration number				
		1	2	3	4	5
100	1	1.03	1.04	1.73	7.8	20.39
	2	1.04	1.02	3.98	14.83	30.61
	3	1.05	1.25	5.39	16.32	32.31
	5	1.08	2.03	10.77	18.82	31.46
1000	1	1.01	1.02	1.54	8.03	18.78
	2	1.01	1.04	3.5	13.84	21.77
	3	1.01	1.26	6.25	16.12	27.57
	5	1.06	2.27	8.27	16.65	35.29
2000	1	1.02	1.02	1.83	8.08	18.33
	2	1.04	1.05	3.53	13.31	25.38
	3	1.01	1.38	6.98	14.39	36.48
	5	1.02	2.29	10.91	22.16	49.89

the gene (Wang *et al.*, 2010; Golightly and Wilkinson, 2005; Reinker *et al.*, 2006).

This gene regulatory network consists of eight chemical reactions which are given below:





Here DNA , P , P_2 and $mRNA$ represent promoter sequences, proteins, protein dimers and messenger RNA respectively. In this network, reactions R_3 and R_7 represent transcription and translation processes in which mRNAs and proteins are synthesized, and then it follows by reactions R_4 and R_8 which are degradation processes. The proteins P and a protein dimer P_2 can be interchanged through reactions R_5 and R_6 under different rate constants. P_2 can be furthermore bound or unbound to DNA through reactions R_1 and R_2 . When a protein dimer binds to the promoter, it represses $mRNA$ production. Overall, the network implements a self-regulatory mechanism to control the synthesis of the protein product, suppressing the transcription when the protein product is abundant (Wang *et al.*, 2010).

We applied the ABC algorithm with initial condition of copy numbers $DNA = 10$, $mRNA = 100$, $P_2 = 800$, $P = 100$, $DNA \cdot P_2 = 100$ and the reaction rate constants of $\mathbf{c} = (0.1, 0.7, 0.35, 0.3, 0.1, 0.9, 0.2, 0.1)$. Similarly, we simulate experimental data for each molecule during a period of $T = 50$ in a step size of $\Delta t = 1$ and results are presented by Fig. 5.4.

The prior distribution we assumed for each estimated parameter follows a uniform distribution $\pi(\theta) \sim U(0, B)$, i.e. for rate constants $c_1 \sim c_8$, the values of B are $(0.5, 2, 1, 0.1, 0.5, 5, 1, 0.1)$. The system is then simulated over five iterations using various step sizes ($\Delta t = 1, 2, 5, 10$) and simulation numbers $(100, 1000, 2000)$.

One example for the simulation results of those eight estimated parameters is presented by Fig. 5.5, which takes a simulation number of 100 and step size of 5. The probability distribution of the estimated rate constant c_7 over iterations ($2 \sim 5$) is shown in Fig. 5.5. Although most of the results for the estimated parameters do not approach the exact rate constant, we still find similar patterns as the

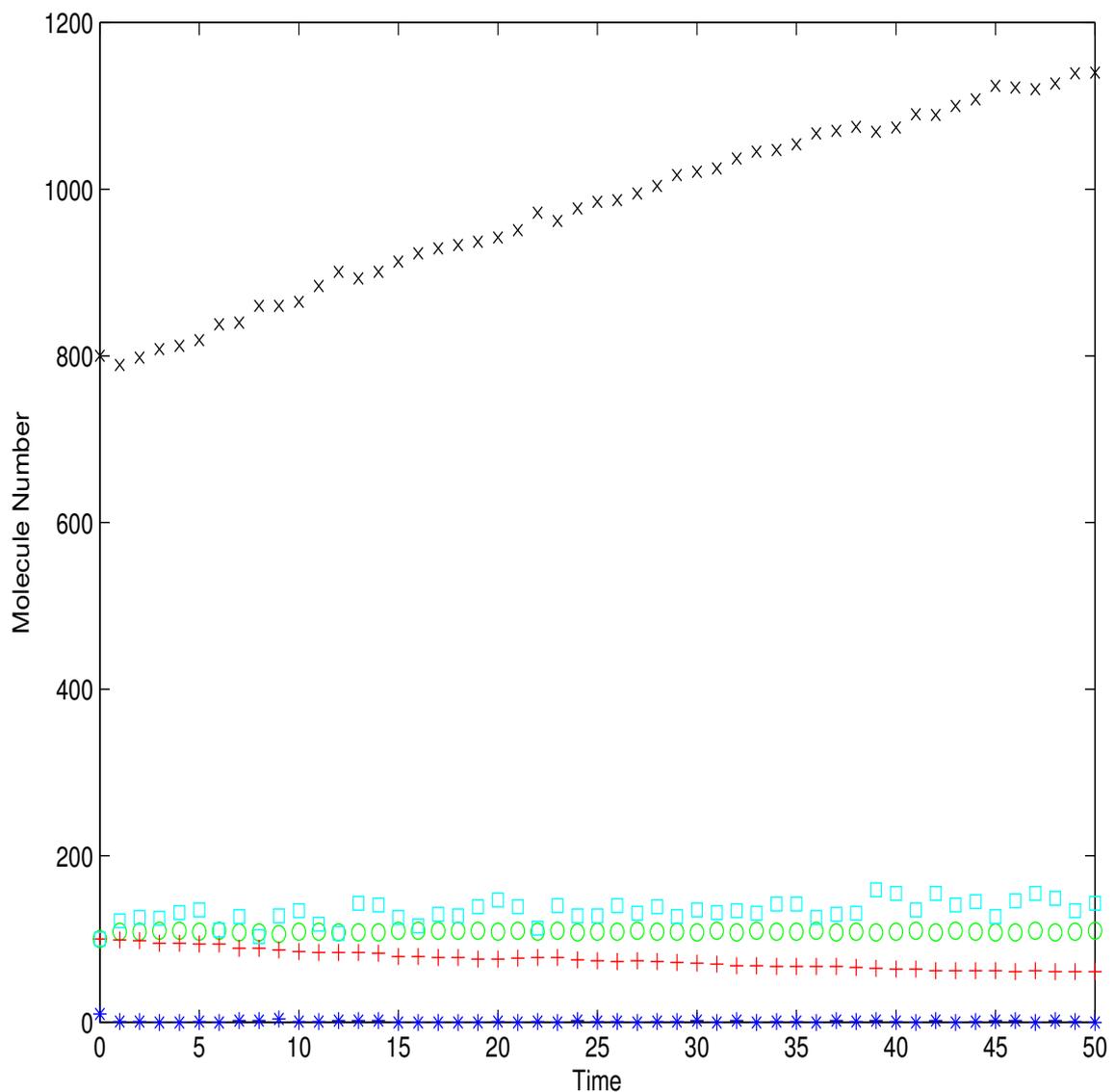


Figure 5.4: Simulated experimental data for system dynamics in a time length of 50 with step size Δt of 1 (Blue star for DNA; green circle for DNA.P₂ and red cross for mRNA black; cyan square for P; black x-mark for P₂).

distribution tends to the centre at the exact rate constants with a normally-like distribution.

For this system, we obtained the mean count as well as the averaged error for each iteration. Fig. 5.6 illustrates an example for the value of mean count number and averaged error with a simulation number of 100 and step size of 1 and 5, respectively. This figure shows a trend that cases with larger step size take more

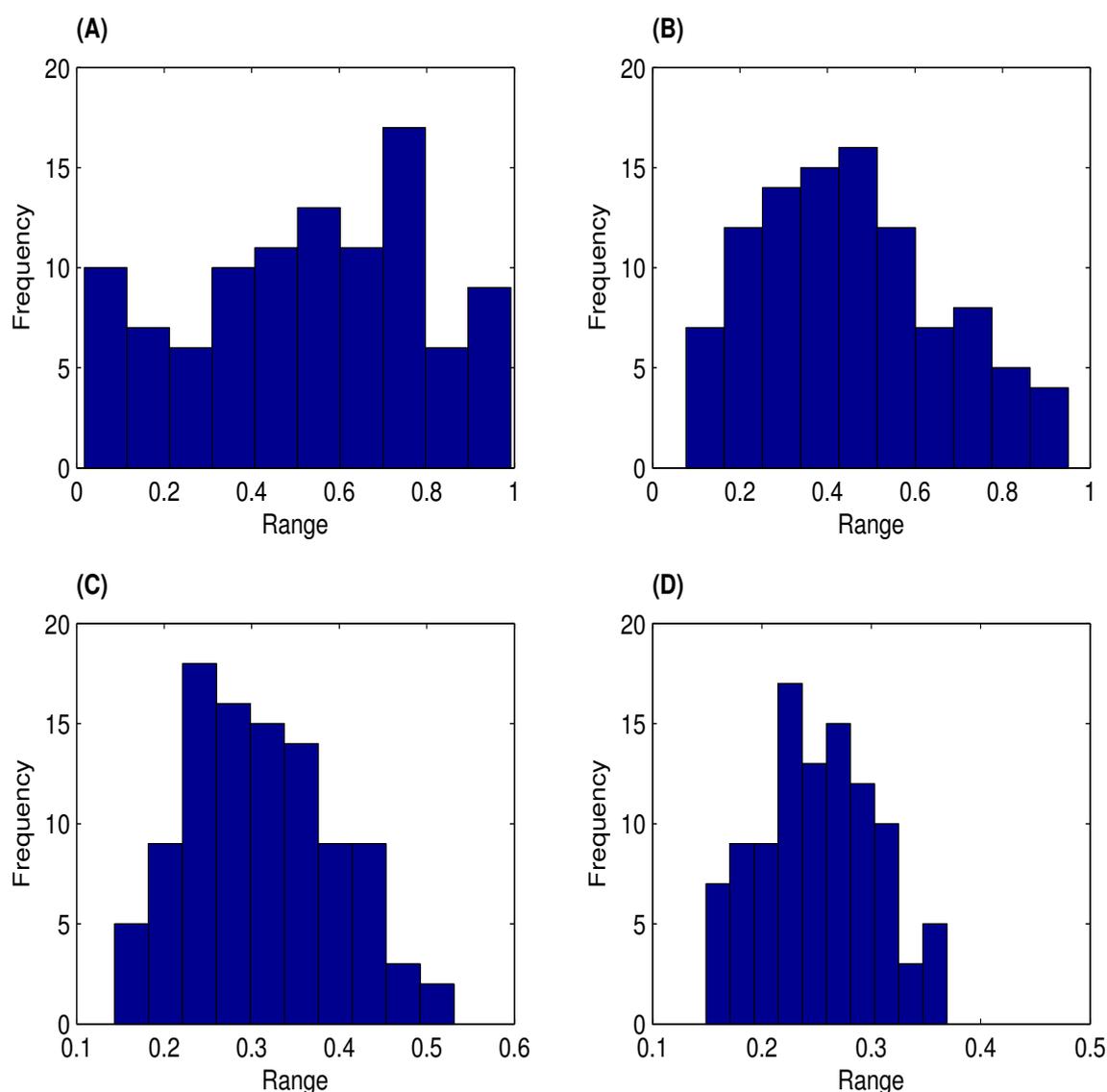


Figure 5.5: Probability distribution of estimated rate constant of c_7 over four iterations ((A):Iteration 2; (B): 3; (C):4; (D):5).

counts to achieve same accuracy with a smaller step size. Comparing with the first system we tested, this system has twice the number of unknown parameters, and the maximum value of mean count for the second system is much larger than that of the first system. In addition, since the molecular numbers in the second system are quite small, the fluctuations in the copy numbers have much influence on the accuracy of the estimated parameters. Fig. 5.6 also suggests that, when the count

number increases over the iterations, the averaged errors of the estimates become smaller, which is consistent with the results in Fig. 5.3.

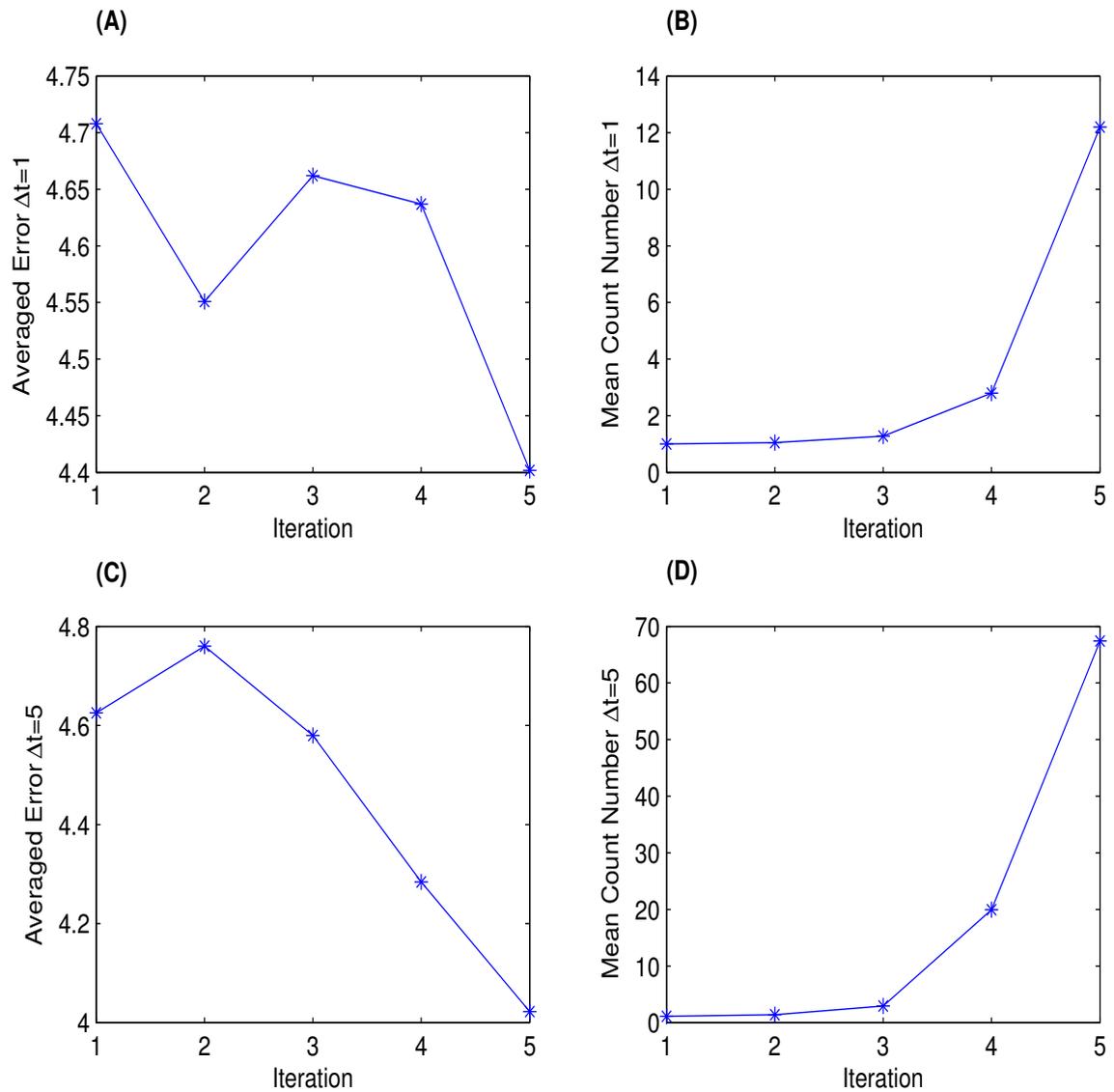


Figure 5.6: The averaged error of estimated parameters and mean count number of iterations with step size Δt of 1 ((A), (B)) and 5 ((C), (D)).

5.4 Conclusion

In this work we conducted extensive computational tests to examine the influence of a number of factors on the estimation error of the ABC algorithm. The ABC

algorithm is an effective inference method that is capable of dealing with inference problems whose likelihood functions are hard to compute. Using two chemical reaction systems as the test problem, we obtained results of the system of four chemical reactions based on the ABC algorithm with various simulation numbers and different step sizes under proper threshold values. From that, we noticed that taking different step sizes would not lead to distinct results. In addition, we examined the influence of a number factors on the performance and accuracy of the ABC algorithm. Our results suggested that the ABC algorithm is a promising method that can be used to infer parameters in high-dimensional and complex biological system models.

Numerical results suggest that a larger count number leads to estimates with better accuracy, and the threshold value determines the count number. Similar to the inference problem for deterministic models, the selection of proper threshold values is a key challenge in the inference for stochastic models. A relatively larger threshold value may generate estimates whose iteration count numbers are always one while a relatively smaller threshold value may lead to estimates with very large count numbers. In addition, a smaller threshold value cannot ensure estimates with better accuracy. Thus more sophisticated techniques, such as the adaptive selection methods, are needed to select the threshold values in the ABC algorithms.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 6

Declaration by candidate

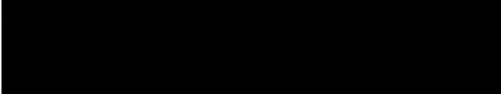
In the case of Chapter 6, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Developed, established and verified the method Wrote programming codes and the article	90%

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Kate Smith-Miles	Provided helpful guidance and proofreading	
Tianhai Tian	Provided helpful guidance and editorial work	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature		Date 29/04/15
Main Supervisor's Signature		Date 29/04/15

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 6

Approximate Bayesian Computation Schemes for Parameter Inference of Discrete Stochastic Models using Simulated Likelihood Density

Chapter 6 is based on the article Wu Q, Smith-Miles K, Tian T. 2014. Approximate bayesian computation schemes for parameter inference of discrete stochastic models using simulated likelihood density. BMC bioinformatics 15(S12): S3, doi: 10.1186/1471-2105-15-S12-S3, URL <http://www.biomedcentral.com/1471-2105/15/S12/S3>.

Abstract.

Background: *Mathematical modelling is an important tool in systems biology to study the dynamic property of complex biological systems. However, one of the major challenges in systems biology is how to infer unknown parameters in mathematical models based on the experimental data sets, in particular, when the data are sparse and the regulatory network is stochastic.*

Results: *To address this issue, this work proposed a new algorithm to estimate parameters in stochastic models using simulated likelihood density in the framework of approximate Bayesian computation. Two stochastic models were used to demonstrate the efficiency and effectiveness of the proposed method. In addition, we designed another algorithm based on a novel objective function to measure the accuracy of stochastic simulations.*

Conclusions: *Simulation results suggest that the usage of simulated likelihood density improves the accuracy of estimates substantially. When the error is measured at each observation time point individually, the estimated parameters have better accuracy than those obtained by a published method in which the error is measured using simulations over the entire observation time period.*

References are considered at the end of the thesis.

Chapter 6

Approximate Bayesian Computation Schemes for Parameter Inference of Discrete Stochastic Models using Simulated Likelihood Density

6.1 Introduction

In recent years, quantitative methods have become increasingly important for studying complex biological systems. To build a mathematical model of a complex system, two main procedures are commonly conducted (Zhan and Yeung, 2011). The first step is to determine the elements of the network and regulatory relationships between the elements. In the second step, we need to infer the model parameters according to experimental data. Since biological experiments are time-consuming and expensive, normally experimental data are often scarce and incomplete compared with the number of unknown model parameters. In addition, the likelihood surfaces of large models are complex. The calibration of

these unknown parameters within a model structure is one of the key issues in systems biology (Kikuchi *et al.*, 2003). The analysis of such dynamical systems therefore requires new, effective and sophisticated inference methods.

During the last decade, several approaches have been developed for estimating unknown parameters: namely, optimization methods and Bayesian inference methods. Aiming at minimizing an objective function, optimization methods start with an initial guess, and then search in a directed manner within the parameter space (Gadkar *et al.*, 2005; Gonzalez *et al.*, 2007). The objective function is usually defined by the discrepancy between the simulated outputs of the model and sets of experimental data. Recently, the objective function has been extended to a continuous approach by considering simulation over the whole time period (Deng and Tian, 2014) and a multi-scale approach by including multiple types of experimental information (Tian and Smith-Miles, 2014). Several types of optimization methods can be found in the literature, among which two major types are called gradient-based optimization methods and evolutionary-based optimization methods. Based on these two basic approaches, various techniques such as simulated annealing (Kirkpatrick *et al.*, 1983), linear and non-linear least-squares fitting (Mendes and Kell, 1998), genetic algorithms (Srinivas and Patnaik, 1994) and evolutionary computation (Ashyraliyev *et al.*, 2008; Moles *et al.*, 2003) have been attempted to build computational biology models. Using optimization methods, the inferred set of parameters produces the best fit between simulations and experimental data (Lall and Voit, 2005; Lillacci and Khammash, 2010), which have been successfully applied for biological systems, however, there are still some limitations with these methods such as the problem of high computational cost when significant noise exists in the system. To address these issues, deterministic and stochastic global optimization methods have been explored (Goel *et al.*, 2008).

When modelling biological systems where molecular species are present in low copy numbers, measurement noise and intrinsic noise play a substantial role (Raj

and van Oudenaarden, 2008), which is a major obstacle for modelling. Bayesian inference methods have been used to tackle such difficulties by extracting useful information from noise data (Wilkinson, 2007). The main advantage of Bayesian inference is that it is able to infer the whole probability distributions of parameters by updating probability estimates using Bayes' Rule, rather than just a point estimate from optimization methods. Also, Bayesian methods are more robust than using other methods when they are applied to estimate stochastic systems, which is not that obvious for modelling of deterministic systems (Toni *et al.*, 2009). Developments have taken place during the last 20 years and recent advances in Bayesian computation including Markov chain Monte Carlo (MCMC) techniques and sequential Monte Carlo (SMC) methods have been successfully applied to biological systems (Battogtokh *et al.*, 2002; Sisson *et al.*, 2007).

For the case of parameter estimation when likelihoods are analytically or computationally intractable, approximate Bayesian computation (ABC) methods have been applied successfully (Beaumont *et al.*, 2002; Marjoram *et al.*, 2003). ABC algorithms provide stable parameter estimates and are also relatively computationally efficient, therefore, they have been treated as substantial techniques for solving inference problems of various types of models that were intractable only a few years ago (Sisson *et al.*, 2007). In ABC, the evaluation of the likelihood is replaced by a simulation-based procedure using the comparison between the observed data and simulated data (Pritchard *et al.*, 1999). Recently, a semi-automatic method has been proposed to construct the summary statistics for ABC (Fearnhead and Prangle, 2012). These methods have been applied in a diverse range of fields such as molecular genetics, epidemiology and evolutionary biology etc. (Marjoram and Tavaré, 2006; Tanaka *et al.*, 2006; Thornton and Andolfatto, 2006).

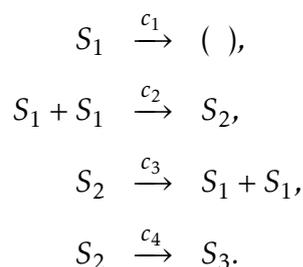
Despite substantial progress in the application of ABC to deterministic models, the development of inference methods for stochastic models is still at the very early stage. Compared with deterministic models, there are a number of open problems

in the inference of stochastic models. For example, recent work proposed ABC to infer unknown parameters in stochastic differential equation models (Picchini, 2014). Our recent computational tests (Wu *et al.*, 2013a) showed the advantages and disadvantages of a published ABC algorithm for stochastic chemical reaction systems in (Toni *et al.*, 2009). In this work, we propose two novel algorithms to improve the performance of ABC algorithms using the simulated likelihood density.

6.2 Results and discussion

6.2.1 The first test system with four reactions

We first examine the accuracy of our proposed methods using a simple model of four chemical reactions (Daigle *et al.*, 2012). The first reaction is the decay of molecule S_1 . Then two molecules S_1 form a dimer S_2 in the second reaction; and this dimerization process is reversible, which is represented by the third reaction. The last reaction in the system is a conversion reaction from molecule S_2 to its product S_3 . All these four reactions are given by



We start with an initial condition with $\mathbf{S} = (10000, 0, 0)$ and rate constants of $\mathbf{c} = (0.1, 0.002, 0.5, 0.04)$, which is termed as the exact rate constants in this test. The stochastic simulation algorithm (SSA) was used to simulate the stochastic

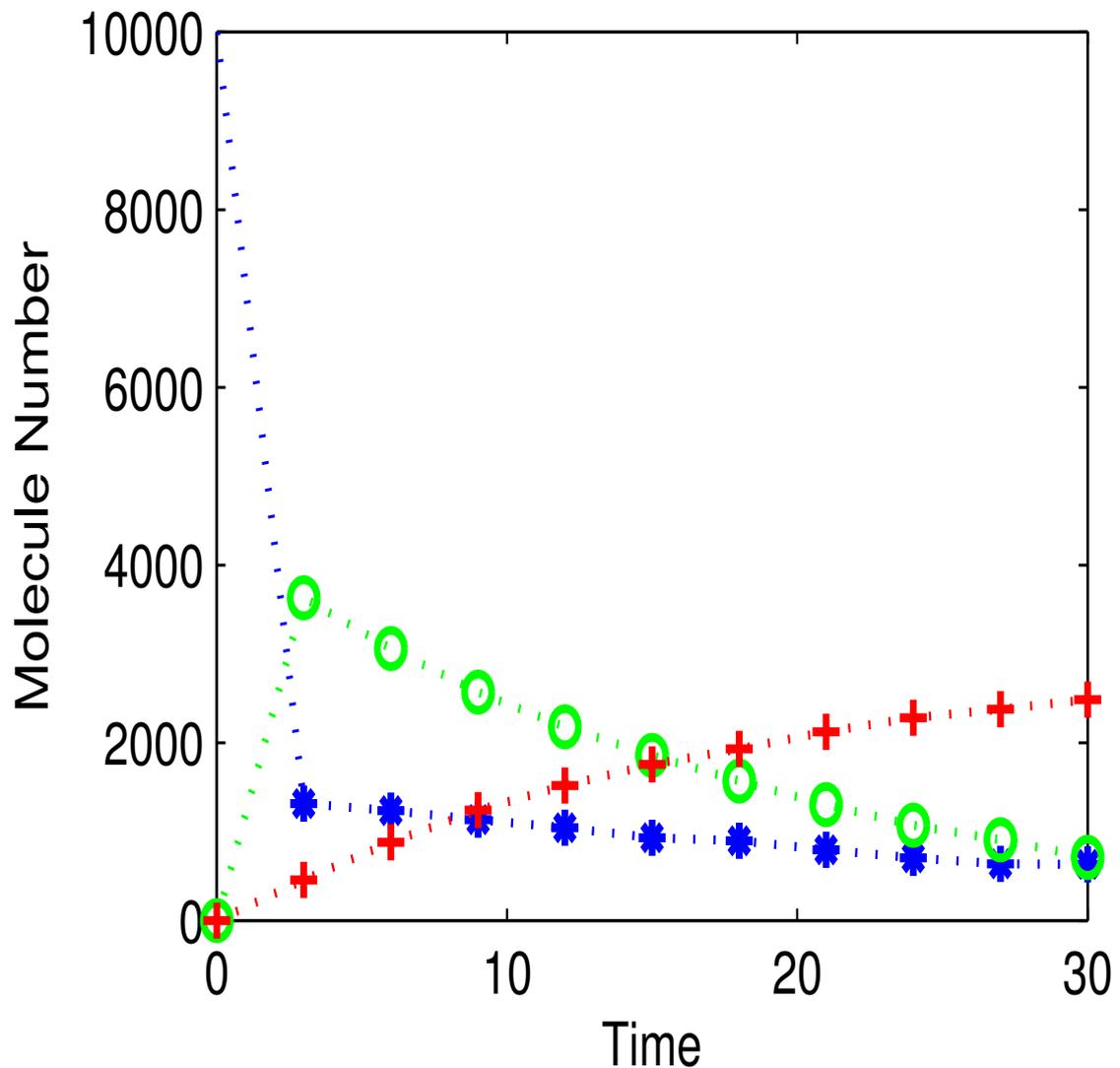


Figure 6.1: Simulated experimental data for system dynamics in a time length of 30 with step size Δt of 3: Blue star for S_1 , green circle for S_2 , and red cross for S_3 .

system (Gillespie, 1977). A single trajectory for this model during a period of $T = 30$ in a step size of $\Delta t = 3$ is presented in Fig. 6.1.

When applying the algorithms in the Method section to estimate model parameters, we assumed the prior distribution for each estimated parameter follows a uniform distribution $\pi(\theta) \sim U(0, A)$. For rate constants $c_1 \sim c_4$, the values of A are (0.5, 0.005, 1, 0.1). Fig. 6.2 shows probabilistic distributions of the estimated rate constant of c_1 over iterations (2 ~ 5). In this test, we have the step size $\Delta t = 3$ and

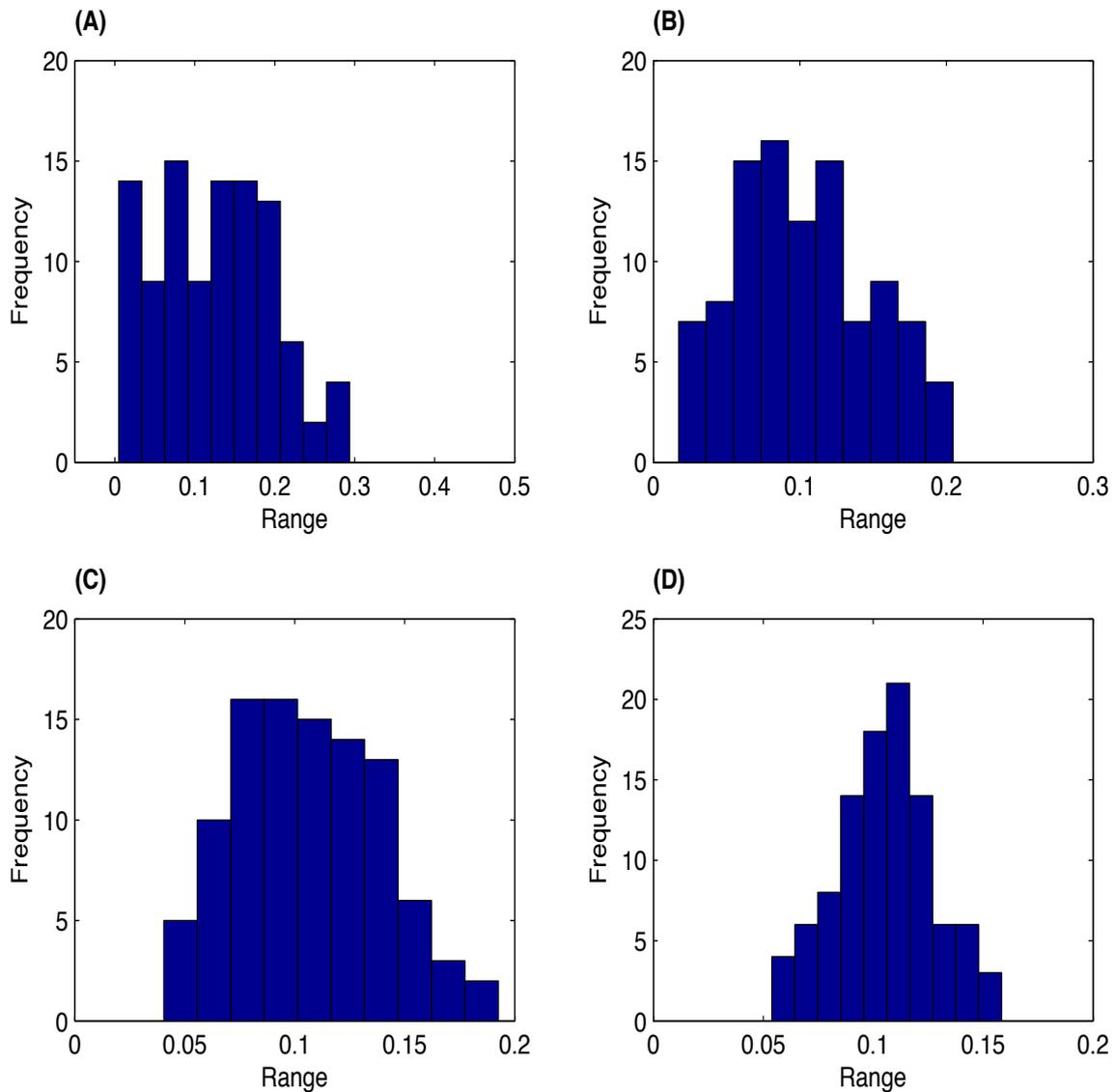


Figure 6.2: Probabilistic distributions of estimated rate constant of c_1 over four iterations using algorithm 1. (A): Iteration 2; (B): 3; (C): 4; (D): 5.

simulation number $B_k = 10$. Figure 6.2 suggests that the probabilistic distribution starts from nearly a uniform distribution in the second iteration (Fig. 6.2(A)) and gradually converges to a normalized-like distribution with a mean value that is close to the exact rate constant.

There are two tolerance values in the proposed algorithms, namely α for the discrepancy in step 2.c and ϵ_k for the fitness error in step 2.d. In the following tests, we considered two strategies: the value of α is a constant (Tian *et al.*, 2007b) or its value varies over iterations. To examine the factors that influence the

convergence rate of particles over iterations, we calculated the mean count number for each iteration, which is the averaged number of counts for accepting all simulated estimation of parameter sets. The averaged error is defined by the sum of relative errors of each rate constant for each iteration. Table 6.1 displays the performances of the tests under three schemes which used fixed discrepancy tolerance $\alpha = 0.1, 0.05$ or varying values of α . In each case, we used the same values of ϵ_k for the fitness tolerance. The value of α in the varying α strategy equals the value of ϵ_k , namely $\alpha_k = \epsilon_k$.

In these performances, we used $\epsilon_k=(0.07, 0.06, 0.055, 0.05, 0.045)$ and $(0.05, 0.045, 0.04, 0.035, 0.03)$ for algorithm 1 with step sizes $\Delta t = 3$ and 5, respectively. For algorithm 2, these values are $\epsilon_k=(0.095, 0.08, 0.065, 0.05, 0.04)$ and $(0.059, 0.055, 0.05, 0.045, 0.04)$. An interesting observation is that the values of mean count number are very large in the first iteration, then decrease sharply and stay within a value stably. We have a detailed test of using different values of the fitness tolerance ϵ_k and found that when using step size of $\Delta t = 3$, mean count number stays at one if $\epsilon_k \geq 0.1$; but it starts to increase sharply to a large number if $\epsilon_k < 0.1$. The observation numbers using a step size of $\Delta t = 3$ is 10 and the maximum error that can incur calculated from step 2.d) is 0.1 with one hundred particles. Similarly, this critical ϵ_k value is 0.06 for a step size of $\Delta t = 5$.

Meanwhile all averaged errors have a decreasing trend over iterations. Looking at different cases with various values of discrepancy tolerance α , it is also observed that using $\alpha = 0.1$ results in more discrepancies of the estimated parameters on average than the other two cases, in particular, than the case $\alpha = 0.05$. Thus in our following tests, we just concentrate on the cases of $\alpha = 0.05$ and varying α . In addition, we observe that by taking $\alpha = 0.05$ for the case with step size of $\Delta t = 3$, it leads to more accurate approximation since $\alpha = 0.05$ is less than most values of α in the case of varying values of α . It is consistent with the cases of a step size of $\Delta t = 5$ in which little differences can be found comparing strategies using

$\alpha = 0.05$ and $\alpha = \epsilon_k$ since the values of ϵ_k are quite close to 0.05. In the case of varying values of α , a small value of ϵ_5 leads to a small value of α_5 , which results in a substantial increase in mean count number. However, this large mean count number does not necessary bring more accurate estimated parameters. With these findings, we simulated results using $\alpha = 0.05$ and $\alpha = \epsilon$ only for algorithm 2. Consistent results are obtained using algorithm 2. Moreover, results obtained using algorithm 2 is more accurate than those from algorithm 1.

6.2.2 The second test system with eight reactions

Although numerical results of the first test system are promising regarding the accuracy, that system has only four reactions. Thus the second test system, namely a prokaryotic auto-regulatory gene network, includes more reactions. This network involves both transcriptional and translational processes of a particular gene. In addition, dimers of the protein suppress its own gene transcription by binding to a regulatory region upstream of the gene (Wang *et al.*, 2010; Golightly and Wilkinson, 2005; Reinker *et al.*, 2006). This gene regulatory network consists of eight chemical reactions which are given below:

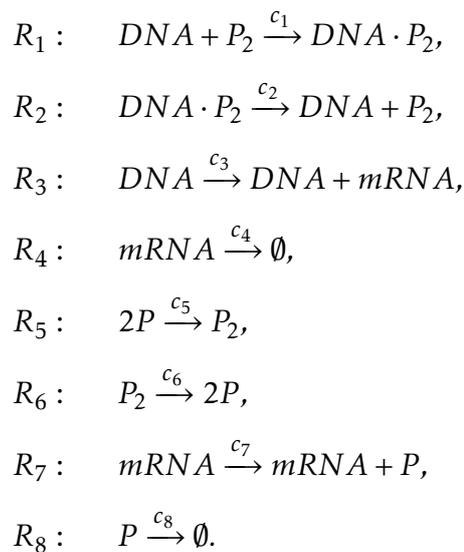


Table 6.1: Comparison of averaged error and mean count number for estimated rate constants over five iterations using algorithms 1 and 2 with simulation number of 10 for system 1. Tests are experimented under different strategies of discrepancy tolerance such as $\alpha = 0.1, 0.05$ or varies over iterations (AE:Averaged Error; MN: Mean count Number).

Δt	$\alpha \backslash k$		1	2	3	4	5	
Algorithm 1								
3	0.1	MN	15.41	7.21	7.36	8.21	10.05	
		AE	0.7668	0.7294	0.7073	0.7832	0.6173	
	0.05	MN	175.72	30.66	24.47	28.22	26.5	
		AE	0.6120	0.5036	0.5521	0.7175	0.6132	
	vary	MN	46.46	25.07	22.76	30.09	88.56	
		AE	0.7669	0.5306	0.6780	0.5858	0.5945	
5	0.1	MN	26.96	10.47	9.07	11.18	13.19	
		AE	0.7107	0.5607	0.5366	0.4693	0.4853	
	0.05	MN	130.64	27.38	25.42	35.36	35.79	
		AE	0.5826	0.6495	0.4260	0.7548	0.4139	
	vary	MN	141.97	30.28	53.47	127.16	2911.58	
		AE	0.5587	0.4793	0.5416	0.5960	0.5375	
Algorithm 2								
3	0.05	MN	467.61	52.34	41.08	69.17	195.69	
		AE	0.5834	0.6091	0.4867	0.4995	0.4402	
	vary	MN	100.26	32.04	24.78	80.15	1793.64	
		AE	0.7132	0.6657	0.6305	0.6705	0.4833	
	5	0.05	MN	333.17	24.26	32.85	21.11	21.84
			AE	0.5962	0.5340	0.5761	0.4983	0.5518
vary		MN	243.78	22.6	31.29	34.6	70.25	
		AE	0.6565	0.6035	0.5759	0.5488	0.4263	

This gene network includes five species, namely DNA, message RNA, protein product, dimeric protein, and the compound formed by dimeric protein binding to the DNA promoter site, which are denoted by DNA, mRNA, P, P_2 and DNA $\cdot P_2$, respectively. In this network, the first two reactions R_1 and R_2 are reversible reactions for dimeric protein binding to the DNA promoter site. Reactions R_3 and R_7 are transcriptional and translation processes for producing mRNA and protein, respectively. Reactions R_5 and R_6 represent the interchange between protein P and dimeric protein P_2 . The system ends up with a degradation process of protein P (Wang *et al.*, 2010).

To apply our algorithms, we start up with an initial condition of molecular copy number

$$(\text{DNA}, \text{mRNA}, \text{P}, P_2, \text{DNA}\cdot P_2) = (10, 100, 100, 800, 100).$$

In addition, the following reaction rate constants

$$(c_1, \dots, c_8) = (0.1, 0.7, 0.35, 0.01, 0.1, 0.9, 0.2, 0.01).$$

are used as the exact rate constants to generate a simulation for each molecular species during a period of $T = 50$ in a step size of $\Delta t = 1$ and results are presented by Fig. 6.3. This simulated dataset is used as observation data for inferring the rate constants.

The prior distribution of each parameter follows a uniform distribution $\pi(\theta) \sim U(0, B)$. For rate constants $c_1 \sim c_8$, the values of B are $(0.5, 2, 1, 0.1, 0.5, 5, 1, 0.1)$. The proposed two algorithms were implemented over five iterations and each iteration contains 100 particles. We choose step sizes $\Delta t = 2$ or 5 and the number of stochastic simulation $B_k = 10$.

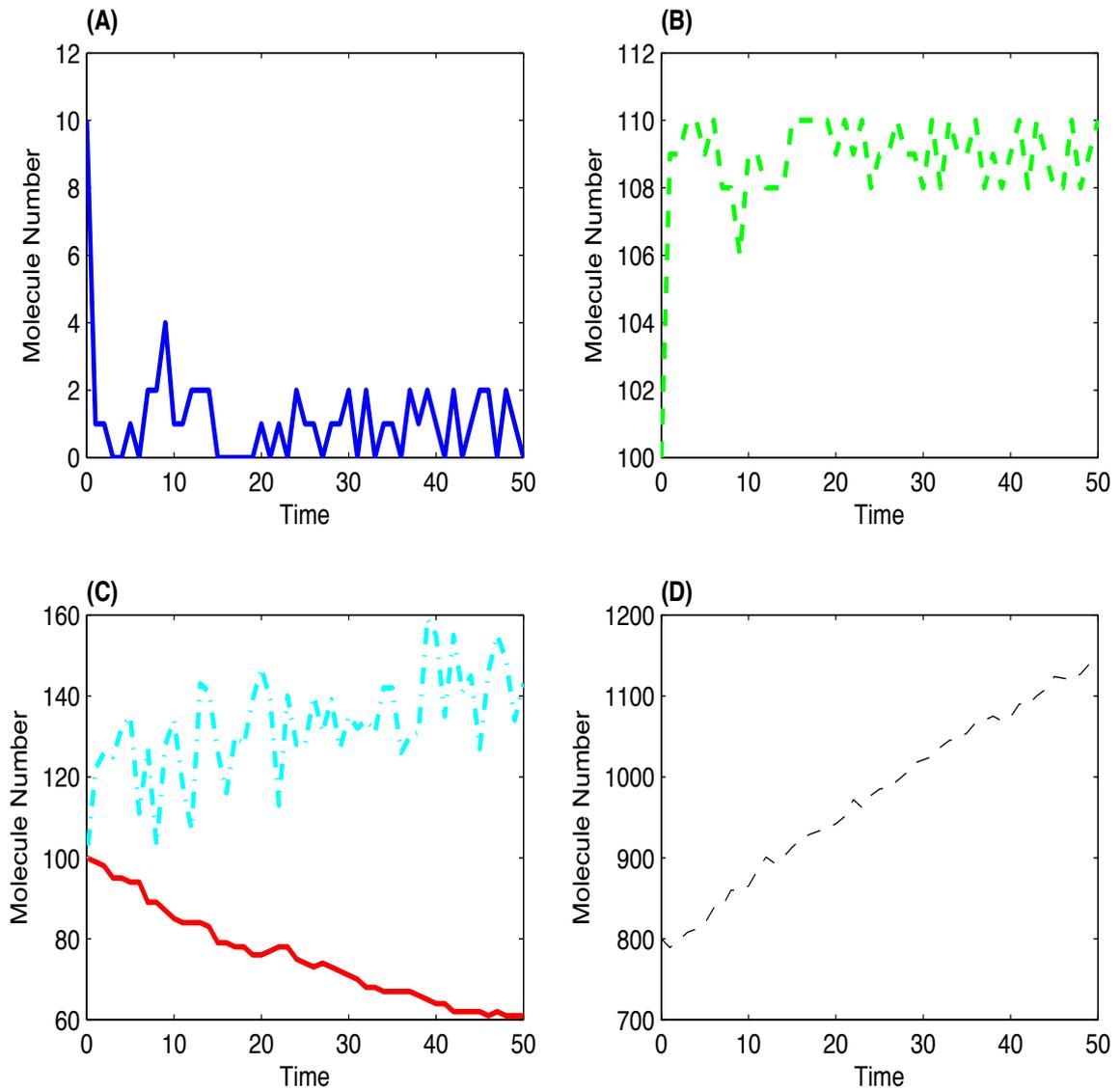


Figure 6.3: Simulated molecular numbers for system 2 in a time length of 50 with step size Δt of 1: (A): DNA numbers; (B): numbers of $\text{DNA}\cdot\text{P}_2$; (C): Red line for the numbers of mRNA black and cyan dash-dotted line for the numbers of P; (D): numbers of P_2 .

Fig. 6.4 gives the probabilistic distribution of the estimated rate constant c_7 over 2nd ~ 5th iterations. The distribution of the first iteration is close to the uniform distribution, and this is not presented. Since the second iteration, the estimated rate constant begins to accumulate around the exact value $c_7 = 0.2$. At the last iteration, the probability in Fig. 6.4 (D) shows a normalized-like distribution. Compared with the results of system 1 in Figure 6.2, the convergence rate of the parameter distribution of system 2 is slower. Our numerical results suggested that this convergence rate depends on the strategy of choosing the values of discrepancy tolerance α .

To analyze the factors that influence the convergence property of estimates, the mean count number as well as the averaged error for each iteration k are obtained. Results are presented in Table 6.2. Using algorithm 1 and 2, we tested for step sizes of $\Delta t = 2$ and $\Delta t = 5$. Since the errors of estimates obtained using a fixed value of $\alpha = 0.1$ are always larger than those obtained by $\alpha = 0.05$, we only tested with the cases of a fixed value $\alpha = 0.05$ and varying values of α . For algorithm 1, we tested two cases for the varying values of discrepancy tolerance α . In the first test, the values are $\epsilon_k = (0.21, 0.2, 0.19, 0.18, 0.175)$ and $\alpha = \epsilon_k$ for varying values of α , which is the case “Same ϵ_k ” in Table 6.2. The values of ϵ_k are also applied to the case of a fixed value $\alpha = 0.05$. In this case, the averaged count number of varying α is much smaller than that of a fixed value of α . Thus we further decreased the value of α to $(0.15, 0.125, 0.1, 0.075, 0.07)$, which is the case “Diff. ϵ_k ” in Table 6.2. In this case, the mean count numbers are similar to those using a fixed α . Numerical results suggested that the strategy of using a fixed value of α generates estimates with better accuracy than the strategies of using varying α values, even when the computing time of the varying α strategy is larger than that of the fixed α strategy.

For algorithm 2, we carried out similar tests. In the first case, we set $\epsilon_k = (0.24, 0.23, 0.22, 0.21, 0.2)$, which is applied to the strategy of fixing $\alpha = 0.05$

Table 6.2: Comparison of averaged error and mean count number for estimated rate constants of system 2 using algorithms 1 and 2. Three strategies are used to choose the discrepancy tolerance α : a fixed value of $\alpha = 0.05$; varying α values; and $\alpha = \epsilon_k$ (denoted as same ϵ_k); varying α values that are smaller than ϵ_k (denoted as diff. ϵ_k). (AE: Averaged Error; MN: Mean count Number).

Δt	$\alpha \backslash k$		1	2	3	4	5
Algorithm 1							
2	0.05	MN	18.29	7.53	9.8	12.7	14.23
		AE	4.6211	4.4179	4.7138	4.2188	3.8119
	Same ϵ_k	MN	2.69	2.07	2.16	1.93	1.93
		AE	4.7006	4.9603	4.8841	4.6833	4.7298
	Diff. ϵ_k	MN	15.26	7.85	8.78	13.06	12.28
		AE	4.8295	4.5322	5.0418	4.7346	4.6069
5	0.05	MN	9.69	3.48	3.12	58.2	74.07
		AE	4.1076	4.3243	4.1868	3.5311	3.5194
	Same ϵ_k	MN	2.34	2.31	2.42	16.9	11.38
		AE	4.9862	4.7669	4.6716	3.8873	4.0017
	Diff. ϵ_k	MN	25.72	8.14	10.45	25.8	174.88
		AE	4.0461	3.9583	3.7474	3.5655	3.6951
Algorithm 2							
2	0.05	MN	89.7	19.75	17.8	40.42	69.52
		AE	4.0540	4.1339	4.1376	3.9696	3.9009
	Same ϵ_k	MN	2.52	3.85	3.55	3.82	3.84
		AE	5.0456	4.6069	4.3666	4.5876	3.8958
	Diff. ϵ_k	MN	197.49	15.05	22.09	36.85	94.24
		AE	3.8712	3.7934	4.3158	3.6485	3.5989
5	0.05	MN	138.14	30.52	46.66	98.87	377.66
		AE	4.0258	3.7218	3.8258	3.8445	3.9205
	Same ϵ_k	MN	21.67	11.34	11.17	26.65	59.64
		AE	4.0545	3.5715	4.1910	3.7252	3.8667
	Diff. ϵ_k	MN	185.54	28.39	33.81	89.81	846.61
		AE	3.7810	3.6694	3.6939	3.9806	3.8515

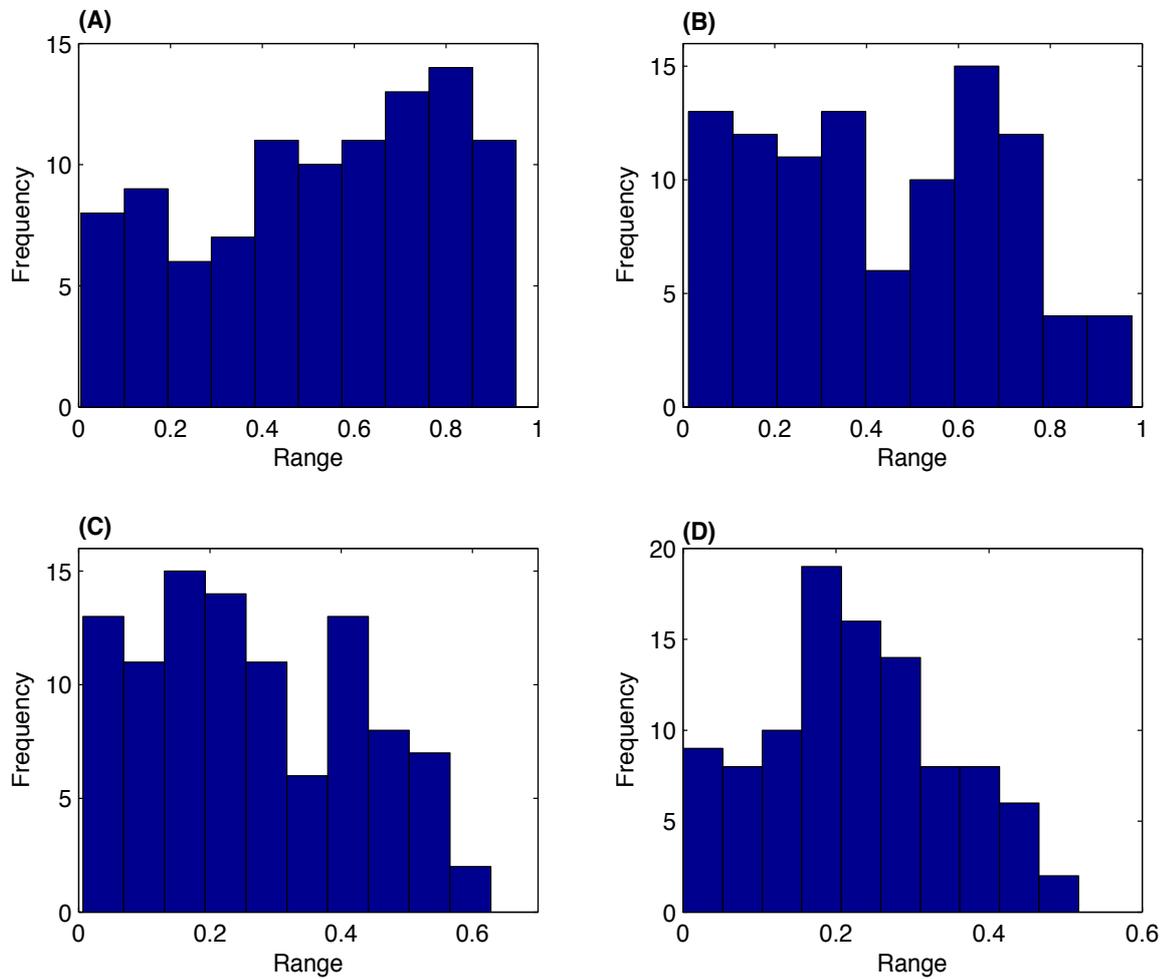


Figure 6.4: Probabilistic distributions of the estimated rate constant c_7 over four iterations using algorithm 1. (A): Iteration 2; (B): 3; (C): 4; (D): 5.

and varying α with $\alpha = \epsilon_k$ that is the case “Same ϵ_k ” in Table 6.2. Again, the averaged count numbers of varying α strategy are much smaller than those using a fixed α . Thus we decreased the value to (0.095, 0.09, 0.085, 0.08, 0.075), which is the case “Diff. ϵ_k ” in Table 6.2; However, the averaged count numbers in the “Diff. ϵ_k ” case are similar to those of the previous two strategies, namely a fixed α and “Same ϵ_k ”. For algorithm 2, Table 6.2 suggests that the varying α strategy generates estimates that are more accurate than those obtained from the fixed α

strategy. However, the best estimates in Table 6.2 are obtained using algorithm 1 and fixed α strategy.

6.3 Conclusion

To uncover the information of biological systems, we proposed two algorithms for the inference of unknown parameters in complex stochastic models for chemical reaction systems. Algorithm 1 is in the framework of ABC SMC and uses transitional density based on the simulations over two consecutive observation time points. Algorithm 2 generates simulations of the whole time interval but differs from the published method in the error finding steps by comparing errors of simulated data to experimental data at each time point. The proposed new algorithms impose stricter criteria to measure the simulation error. Using two chemical reaction systems as the test problems, we examined the accuracy and efficiency of proposed new algorithms. Based on the results of two algorithms for system 1, we discovered that taking smaller values of discrepancy tolerance α will result in more accurate estimates of unknown model parameters. This conclusion is confirmed by the second system that we tested under different conditions. Numerical results suggested that the proposed new algorithms are promising methods to infer parameters in high-dimensional and complex biological system models and have better accuracy compared with the results of the published method (Wu *et al.*, 2013a). The encouraging result is that new algorithms do not need more computing time to achieve such accuracy. Our computational tests showed that the selection for the value of fitness tolerance is a key step in the success of ABC algorithms. The advantage of the population Monte-Carlo methods is the ability to reduce the fitness tolerance gradually over populations. Generally, a smaller value of fitness tolerance will lead to a larger number of iteration count and consequently larger computing time. For deterministic inference problems,

a smaller value of fitness tolerance normally will generate estimates with better accuracy. However, for stochastic models, this conclusion is not always true. In addition to the fitness tolerance, our numerical results suggested that other factors, such as the simulation algorithm for chemical reaction systems and the strategy of discrepancy tolerance, also have influences on the accuracy of estimates. Thus more skilled approaches, such as the adaptive selection process for the fitness tolerance, should be considered to improve the performance of ABC algorithms.

In this work, we used the SSA to simulate chemical reaction systems (Gillespie, 1977). This approach may be appropriate when the biological system is not large. In fact, for the two biological systems discussed in this work, the computing time of inference is still very large. To reduce the computing time, more effective methods should be used to simulate the biological systems, such as the τ -leap methods (Tian and Burrage, 2004) and multi-scale simulation methods (Pahle, 2009; Burrage *et al.*, 2004). Another alternative approach is to use parallel computing to reduce the heavy computing loads. All these issues are potential topics for future research work.

6.4 Methods

6.4.1 ABC SMC algorithm

ABC algorithms bypass the requirement for evaluating likelihood functions directly in order to obtain the posterior distributions of unknown parameters. Instead, ABC methods simulate the model with given parameters, compare the observed and simulated data, and then accept or reject the particular parameters based on the error of simulation data. Thus there are three key steps in the implementations of ABC algorithms. The first step is the generation of a sample of

parameters θ^* from the prior distribution of parameters or from other distributions that are determined in ABC algorithms. The second step is to define distance function $d(\mathbf{X}, \mathbf{Y})$ between the simulated data \mathbf{X} and experimental observation data \mathbf{Y} . Finally, a tolerance value is needed as a selection criterion to accept or reject the sampled parameter θ^* . Based on the generic form of ABC algorithm (Toni *et al.*, 2009), a number of methods have been developed including ABC rejection sampler and ABC MCMC (Boys *et al.*, 2008; Golightly and Wilkinson, 2011). The ABC rejection algorithm is one of the basic ABC algorithm that may result in long computing time when a badly prior distribution that is far away from posterior distribution is chosen. ABC MCMC introduces a concept of acceptance probability during the decision making step which saves computing time. However, this may result in getting stuck in the regions of low probability for the chain and we may never be able to get a good approximation. To tackle these challenges, the idea of particle filtering has been introduced. Instead of having one parameter vector at a time, we sample from a pool of parameter sets simultaneously and treat each parameter vector as a particle. The algorithm starts from sampling a pool of N particles for parameter vector θ through prior distribution $\pi(\theta)$. The sampled particle candidates $(\theta_1^*, \dots, \theta_N^*)$ will be chosen randomly from the pool and we will assign each particle a corresponding weight w to be considered as the sampling probability. A perturbation and filtering process following through a transition kernel $q(\cdot|\theta^*)$ finds the particles θ^{**} . Similarly with θ^{**} , data \mathbf{Y} can be simulated and compared with experimental data \mathbf{X} to further fulfil the requirements for estimating posterior distribution.

The basic form of algorithm described above is as follows (Sisson *et al.*, 2007):

Algorithm: ABC SMC

1. Define the threshold values $\epsilon_1, \dots, \epsilon_K$, start with iteration $k = 1$.

2. Set the particle indicator $i = 1$.
3. If $k = 1$, sample θ^* from the proposed prior distribution $\pi(\theta)$. Generate a candidate data set $D_{(b)}(\theta^*)$ B_k times and calculate the value of $b_k(\theta^*)$, where $D_{(b)} \sim p(D|\theta)$ for any fixed parameter θ ,

$$b_k(\theta^*) = \sum_{b=1}^{B_k} \mathbf{1}(d(D_0, D_{(b)}(\theta^*)) \leq \epsilon_k) \quad (6.4.1)$$

and D_0 is the experimental data set.

If $k > 1$, sample θ from the previous population $\{\theta_{k-1}^i\}$ with weights w_{k-1} and perturb the particle to obtain θ^* using a kernel function \mathbb{K}_k .

If $\pi(\theta^*) = 0$ or $b_k(\theta^*) = 0$, return to the beginning of step 3.

4. Set $\theta_k^i = \theta^*$ and determine the weight for each estimated particles θ_k^i ,

$$w_k^{(i)} = \begin{cases} b_k(\theta_k^i) & \text{if } k = 1; \\ \frac{\pi(\theta_k^i) b_k(\theta_k^i)}{\sum_{j=1}^N \mathbb{K}_k(\theta_{k-1}^j, \theta_k^i)} & \text{if } k > 1. \end{cases}$$

If $i < N$, update $i = i + 1$ and return to step 3.

5. Normalize the weights $w_k^{(i)}$. If $k < K$, update $k = k + 1$ and go back to step 2.

A number of algorithms have been developed using the particle filtering technique, such as the partial rejection control, population Monte-Carlo and SMC. Each of them differs in the formation of weight w and the transition kernels.

6.4.2 ABC using simulated likelihood density

ABC SMC method uses the simulation over the entire time period to measure the fitness to experimental data, which is consistent to the approaches used for deterministic models (Toni *et al.*, 2009). For stochastic models, the widely

used approach is treating transitional density as the likelihood function (Hurn *et al.*, 2007; Hurn and Lindsay, 1999). Based on a sequence of $n + 1$ observations $\mathbf{X} = [X_0, X_1, \dots, X_n]$ at time points $[t_0, t_1, \dots, t_n]$, for a given parameter set θ , the joint transitional density is defined as

$$f_0[(t_0, X_0)|\theta] \prod_{i=1}^n f[(t_i, X_i)|(t_{i-1}, X_{i-1}), \dots, (t_0, X_0); \theta], \quad (6.4.2)$$

where $f_0[\cdot]$ is the density of initial state, and

$$f[(t_i, X_i)|(t_{i-1}, X_{i-1}), \dots, (t_0, X_0); \theta] \quad (6.4.3)$$

is the transitional density starting from (t_{i-1}, X_{i-1}) and evolving to (t_i, X_i) . When the process X is Markov, the density (6.4.3) is simplified as

$$f[(t_i, X_i)|(t_{i-1}, X_{i-1}); \theta]. \quad (6.4.4)$$

In the simulated likelihood density (SLD) methods, this transitional density is approximated by that obtained from a large number of simulations.

Based on the discrete nature of biochemical reactions with low molecular numbers, it was proposed to use the frequency distribution of simulated molecular numbers to calculate the transitional density (Tian *et al.*, 2007b). The frequency distribution is evaluated by

$$F[X = X_l] = \frac{1}{B_k} \sum_{m=1}^{B_k} [1 - \delta(X_l, X_{ml})]$$

using B_k simulations with the simulated state X_{ml} . Here the function $\delta(x)$ is defined by

$$\delta(X_l, X_{ml}) = \begin{cases} 0 & \text{if } d(X_l, X_{ml}) < \alpha X_l; \\ 1 & \text{else,} \end{cases}$$

where $d(x, y)$ is a distance measure between x and y .

Here we propose a new algorithm that uses the simulated transitional density function as the objective function. Unlike ABC SMC algorithm (Toni *et al.*, 2009), the new method considers the transitional density function from t_{i-1} to t_i only at each step. Based on the framework of ABC SMC, the new algorithm using transitional density is proposed as follows.

ABC SLD algorithm 1

1. Given data \mathbf{X} and any assumed prior distribution $\pi(\theta)$, define a set of threshold values $\epsilon_1, \dots, \epsilon_K$.
2. For iteration $k = 1$,
 - (a) Set the particle indicator $i = 1$, sample $\theta^* \sim \pi(\theta)$.
 - (b) For time step $l = 1, 2, \dots, n$, use initial condition \mathbf{X}_{l-1} and parameter θ^* to generate data \mathbf{Y} at t_l for B_k times.
 - (c) For $m = 1, \dots, B_k$, calculate the value of discrepancy and test for

$$d(\mathbf{X}_l, \mathbf{Y}_{ml}) \leq \alpha \mathbf{X}_l, \quad (6.4.5)$$

where α is a defined constant.

If it is true, let $\beta_{ml}(\theta^*) = 0$, otherwise it is one. Then determine

$$b_l(\theta^*) = \sum_{m=1}^{B_k} \beta_{ml}(\theta^*). \quad (6.4.6)$$

- (d) Calculate

$$\epsilon = \sum_{l=1}^m \frac{1}{B_k} (B_k - b_l(\theta^*)). \quad (6.4.7)$$

If $\epsilon < \epsilon_k$, update $\theta_i^k = \theta^*$ and move to the next particle $i = i + 1$.

- (e) Assign weight $w_i^k = \frac{1}{N}$ for each particle.

3. Determine the variance for the particles in the first iteration

$$\sigma_1 = \sqrt{\text{var}(\theta_{\{1:N\}}^1)}$$

4. For iteration $k = 2, \dots, K$

- (a) Start with $i = 1$, Sample $\theta^* \sim \theta_{i:N}^{k-1}$ using the calculated weights $w_{i:N}^{k-1}$.
- (b) Perturb θ^* through sampling $\theta^{**} \sim q(\theta|\theta^*)$, where $q = N(\theta^*, \sigma_{k-1}^2)$ or $q = U(a, b)$. Here values of a, b depend on θ^* and σ_{k-1}^2 .
- (c) Generate simulations and calculate the error ϵ using the same steps as in 2(b)~(d).
- (d) For each particle, assign weights

$$w_i^k = \frac{\pi(\theta_i^k) b_k(\theta_k^i)}{\sum_{j=1}^N w_j^{k-1} q(\theta_j^{k-1} | \theta_j^k, \sigma_{k-1})}$$

- (e) Determine the variance for the particles in the k -th iteration

$$\sigma_k = \sqrt{\text{var}(\theta_{\{1:N\}}^k)}$$

An alternative approach is to generate simulations over the observation time period but compare the error to experimental data at each time point. The approach locates somewhere between ABC SMC algorithm (Toni *et al.*, 2009) and the proposed Algorithm 1, which is presented below. For simplicity we do not give a detailed algorithm, but just provide the key steps 2.b) ~ 2.d) that are different from those in Algorithm 1.

ABC SLD algorithm 2

2.b) Generate data \mathbf{Y} B_k times using θ^* .

2.c) For $m = 1, \dots, B_k$ and $l = 1, 2, \dots, n$, calculate the value of discrepancy $d(\mathbf{X}_l, \mathbf{Y}_{ml})$ and test for

$$|\mathbf{X}_l - \mathbf{Y}_{ml}| \leq \alpha \mathbf{X}_l.$$

If it is true, let $b_{ml}(\theta^*) = 0$, otherwise it is one.

2.d) Calculate

$$\epsilon = \sum_{l=1}^n \frac{1}{B_k} \sum_{m=1}^{B_k} b_{ml}(\theta^*).$$

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 7

Declaration by candidate

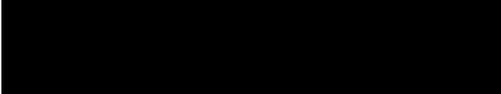
In the case of Chapter 7, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Developed, established and verified the method Wrote programming codes and the article	90%

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Feng Jiang	Provided helpful guidance and proofreading	
Tianhai Tian	Provided helpful guidance and editorial work	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature		Date 29/04/15
Main Supervisor's Signature		Date 29/04/15

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 7

Sensitivity and Robustness Analysis for Stochastic Model of Nanog Gene Regulatory Network

Chapter 7 is based on the article Wu Q, Jiang F, Tian T. 2015. Sensitivity and robustness analysis for stochastic model of nanog gene regulatory network. International Journal of Bifurcation and Chaos 25(07): 1540 009, doi: 10.1142/S021812741540009X, URL <http://www.worldscientific.com/doi/abs/10.1142/S021812741540009X>.

Abstract. *The advances of systems biology have raised a large number of mathematical models for exploring the dynamic property of biological systems. A challenging issue in mathematical modelling is how to study the influence of parameter variation on system property. Robustness and sensitivity are two major measurements to describe the dynamic property of a system against the variation of model parameters. For stochastic models of discrete chemical reaction systems, although these two properties have been studied separately, no work has been done so far to investigate these two properties together. In this work, we propose an integrated framework to study these two properties for a biological system simultaneously. We also consider a stochastic model with intrinsic noise for the Nanog gene network based on a published model that studies extrinsic noise only. For the stochastic model of Nanog gene network, we identify key coefficients that have more impacts on the network dynamics than the others through sensitivity analysis. In addition, robustness analysis suggests that the model parameters can be classified into four types regarding the bistability property of Nanog expression levels. Numerical results suggest that the proposed framework is an efficient approach to study the sensitivity and robustness properties of biological network models.*

Keywords. *Genetic regulatory network; sensitivity analysis; robustness property; stochastic model; simulation.*

References are considered at the end of the thesis.

Chapter 7

Sensitivity and Robustness Analysis for Stochastic Model of Nanog Gene Regulatory Network

7.1 Introduction

Recent advances in systems biology have demonstrated that mathematical modelling is a very important role to study the dynamic property of biological networks at system level. A key step in model development is to determine the values of unknown model parameters. Since unknown parameters are difficult and sometimes even impossible to measure with biological experiments, a number of numerical methods have been developed recently. These numerical methods can mainly classified as optimization approaches or Bayesian approaches methods (Wu *et al.*, 2014; Hartig *et al.*, 2011; Tian *et al.*, 2007b). Despite these progresses, the inference of model parameters still remains as a challenging issue in system biology. For example, some estimated parameters may have values with large variations in different numerical tests, and these sets of parameters all can faithfully realize

experimentally observed data. Thus, our confidence on the model predictions may be limited due to the uncertainties of model parameters.

To address these challenges, sensitivity analysis is a major technique to determine how fluctuations in the output of mathematical models can be influenced by the changes in the model inputs (Gunawan *et al.*, 2005). For example, the most sensitive model parameters and their corresponding biological processes may be the potential targets for further experimental analysis for drug designs. In recent years, sensitivity analysis has become an increasingly important step in mathematical modelling (Marino *et al.*, 2008). It may also provide important measures to the model parameters in the parameter inference step of model development (Kiparissides *et al.*, 2009). In general, methods for sensitivity analysis can be classified into two major approaches: local and global methods. Local methods study the influence of a single parameter in isolation and the other parameters are kept constant at their nominal values. On the other hand, global methods study the influence of a parameter by varying it in a defined direction and also simultaneously varying the other parameters in a random fashion in the entire parameter space (Saltelli *et al.*, 2008). For example, the method of derivative based global sensitivity measures (DGSM) has recently become popular among practitioners (Kucherenko and Iooss, 2014). However, the main drawback of the global methods is their extensive computational costs for large models.

Chemical reaction model has been regarded as a popular approach in the last ten years to investigate both intrinsic noise and extrinsic noise in biological networks. These models typically depend on a set of kinetic parameters whose values are often unknown or fluctuate due to an uncertain environment (extrinsic noise). For gene regulatory network models, small changes to the parameters may significantly alter the system output, and thus it is critical to characterize such effects. Although parametric sensitivity analysis is an indispensable analysis technique in the study of kinetic models, classical sensitivity analysis does not directly apply

to discrete stochastic dynamical systems, which has recently gained popularity because of its relevance in the simulation of biological processes. In the stochastic setting, the simplest and most common method for finite perturbations is to compute a finite difference via Monte Carlo simulations (Gunawan *et al.*, 2005; Rathinam *et al.*, 2010). Sensitivity analysis for discrete stochastic processes has been developed based on density function (distribution) sensitivity using an analog of the classical sensitivity and the Fisher Information Matrix (Gunawan *et al.*, 2005). Recently this algorithm has been formulated in (Damiani *et al.*, 2013) and applied to study the sensitivity property of catalytic reaction networks.

Robustness, in both biological and engineering systems, can be defined as the ability of a system to function correctly in the presence of both internal and external uncertainties (Bates and Cosentino, 2011). It was firstly introduced by Csete and Doyle (2002), which then has been extensively studied by Kitano and co-workers (Kitano, 2004, 2007). Since robustness is an ubiquitously observed property of biological systems (Kitano, 2004; Tian *et al.*, 2011), this property has been widely used recently as an important measure to select the optimal network structure or model rate constants from estimated candidates (Citri and Yarden, 2006; Apri *et al.*, 2010; Masel and Siegal, 2009). A formal and abstract definition of the robustness property, given by Kitano (2007), has been widely used in analyzing robustness properties of biological systems (Tian and Song, 2012).

Sensitivity property and robustness represent two different aspects of dynamic properties of biological network models, namely the variation of output and maintenance of system property, respectively. This work will use a gene network as the test system to investigate these two property simultaneously. Embryonic stem cells (ESC) have the ability to self-renew and remain pluripotent, while continuously providing a source of variety for differentiated cell types. Understanding the regulatory mechanisms for controlling these properties at molecular level is very important for stem cell biology and its application to regenerative medicine.

Studying the dynamic property at a system level is crucial for elucidating those molecular interactions, which regulate the reprogramming of somatic cells into ESC (Chickarmane *et al.*, 2012). The maintenance of the pluripotent state of ESC over a number of self-renewing divisions is associated with a characteristic expression pattern of a number of particular genes. Extensive experimental studies have demonstrated that the transcription factors (TF) Oct4, Sox2 and Nanog play an important role in this regulatory process by directing the gene expression in ESC through a cooperative interaction (Rodda *et al.*, 2005). Biological experiments in recent years also suggest that ESC populations are heterogeneous with respect to the expression levels of Nanog and that individual ESCs reversibly change their Nanog expression levels (Glauche *et al.*, 2010; Navarro *et al.*, 2012). In addition, similar expression patterns have been found for TF Rex1, which is a reliable marker for undifferentiated ESC and described as a downstream target of TFs Nanog, Oct4 and Sox2 (Toyooka *et al.*, 2008). Furthermore, Autocrine FGF4/Erk signalling has been identified as a major stimulus for fate decisions and lineage commitment in ESC (Kunath *et al.*, 2007). Taken together, these genes and external stimulus are the critical determinants for the degree of ESC heterogeneity and differentiation.

Mathematical modelling have been designed as a powerful tool to explore the function of regulatory mechanisms in the Nanog gene network. The first mathematical model of the Oct4-Sox2-NANOG network motif was developed for studying the bistable switching due to several positive feedback loops and environmental signals (Chickarmane *et al.*, 2006). A subsequent stochastic model was designed to explore mechanisms and feedback regulations and describe the observed variation of the Nanog levels in mouse ESC (Glauche *et al.*, 2010). In addition, a more detailed computational model has been proposed; and stochastic simulations suggest that NANOG heterogeneity is the deciding factor for the stem cell fate

(Chickarmane *et al.*, 2012). Recently, a new mathematical model has been proposed to explicitly integrate FGF4/Erk signalling into an interaction network of key pluripotency factors, namely Oct4, Sox2, Nanog and Rex1. Simulation results suggest that interactions between FGF4/Erk signalling and Nanog expression qualify as the major mechanism to manipulate mouse ESC pluripotency (Herberg *et al.*, 2014).

Although the function of extrinsic noise has been studied, there has not been any work so far to explore the influence of intrinsic noise on the dynamics of Nanog gene network. In this work, we first propose a general framework to study sensitivity and robustness properties simultaneously based on stochastic simulations. A stochastic model of the Nanog gene network is developed in order to explore the function of internal noise in the system state. Using our proposed framework, we study the sensitivity and robustness properties of this stochastic model regarding the variation of each parameter. The remaining part of this paper is organized as follows. Section 7.2 proposes the computational framework and develops the stochastic model of the Nanog gene network. Section 7.3 presents simulation results of the stochastic model, and as well as the sensitivity and robustness properties of the model. Discussion of the connection and difference between sensitivity and robustness is presented in Section 7.4.

7.2 Method

In this section we first derive the stochastic model using chemical equations based on the developed stochastic differential equation model (Herberg *et al.*, 2014). Then we propose a general framework to calculate the sensitivity coefficient and robustness property of discrete models for this chemical reaction system.

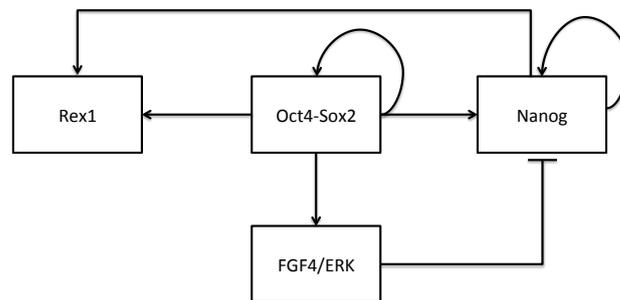
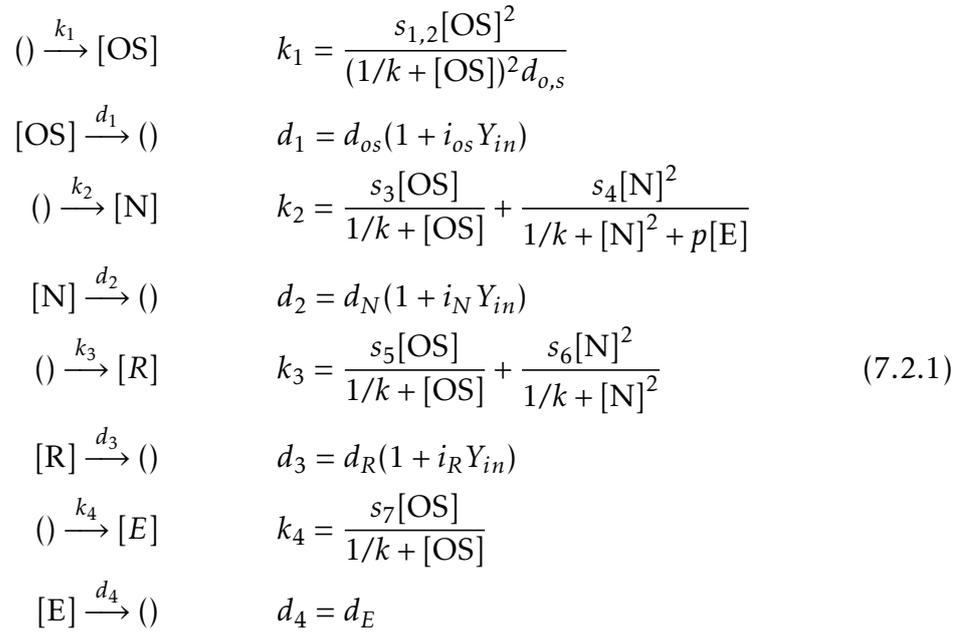


Figure 7.1: Network diagram for the Nanog gene regulatory network

7.2.1 Mathematical model

A mathematical model has been developed recently to study the function of external noise in the genetic regulation of the Nanog network (Herberg *et al.*, 2014). This network considers the TFs Oct4, Sox2, Nanog and Rex1 as central elements of a self-regulating intracellular network structure. It is assumed that Oct4 and Sox2 proteins form heterodimers to positively regulate their own expressions and to activate the transcription of Nanog and Rex1 to a basal level (Shi *et al.*, 2006). A Hill-coefficient $n = 2$ is selected in the mathematical formulation according to the finding that Nanog proteins form homodimers (Wang *et al.*, 2008). In addition to the basal activation of Rex1 through Oct4 and Sox2, Nanog is considered to be an activator for the transcription of the pluripotency marker Rex1 (Shi *et al.*, 2006). Moreover, there are experimental evidences showing that Oct4 and Sox2 induce Erk activity through the activation of FGF4 and that Erk signalling acts as a potential Nanog repressor (Silva *et al.*, 2009). Hence, a negative FGF4/Erk-mediated feedback loop is included in this stochastic network model (see Fig 7.1).

For the quantitative assessment of the model structure, a mathematical description has been derived for the interactions between the TFs Oct4, Sox2, Nanog, Rex1, and the signalling pathways FGF4/Erk (Herberg *et al.*, 2014). Based on this model, we derive a stochastic model using chemical reactions. Here [OS], [N], [R] and [E] are denoted as the protein copy numbers of Oct4-Sox2, Nanog, Rex1 and FGF4/Erk, respectively. The temporal changes of the protein concentrations [OS], [N], [R] and [E] are represented by the following set of discrete chemical reactions:



where p is the repression rate to the expression of gene Nanog, which is regulated by the FGF4/Erk pathway. All proteins and protein complexes are degraded by first-order kinetics with protein specific degradation rates d_j (with $j \in (\text{OS}, \text{N}, \text{R}, \text{E})$). The degradation rate are enhanced by inhibition factors i_j depending on the intracellular activity of a differentiation signal Y , denoted by Y_{in} . However, Y_{in} is not considered in this work. Thus we assume $Y_{in} = 0$. There are two sets of rate constants for the above model (Herberg *et al.*, 2014). Here we apply the first set that has the following values: $s_{1,2} = 75$, $s_3 = 0.1$, $s_4 = 40$, $s_5 = 15$, $s_6 = 140$, $s_7 = 2$, $k = 0.1$, $k_\gamma = 4$, $i_{OS} = 0$, $i_N = 0$, $i_R = 0$, $d_{O,S} = 0.01$, $d_{OS} = 1$, $d_N = 1$, $d_R = 1$, $d_S = 1$.

7.2.2 Framework for sensitivity and robustness analysis

Gunawan *et al.* (2005) proposed the first numerical method to calculate the sensitivity property of a discrete stochastic model of chemical reaction systems. This method aims at calculating the sensitivity measure, given by

$$S_{X,\theta}(t) = E \left[\left| \frac{\partial f(X,\theta,t)}{\partial \theta} \right| \right] = \int_{\Omega_\theta} \int_{\Omega_X} \left| \frac{\partial f(X,\theta,t)}{\partial \theta} \right| f(X,\theta,t) d_\theta d_X \quad (7.2.2)$$

where Ω_θ and Ω_X are the domains of integration of the parameter θ and variable X , respectively. Here function $f(X,\theta,t)$ is the density function of variable X with parameter θ at time point t . Thus in order to calculate this sensitivity measure, we not only need to estimate the density function based on stochastic simulations but also determine the partial derivative of the density function.

Regarding robustness analysis, a formal and abstract definition (Kitano, 2007) is used in this work to measure the robustness property of the proposed model. The robustness property of a mathematical model with respect to a set of perturbations P is defined as the average of an evaluation function $D_{a,P}^s$ of the system over all perturbations $p \in P$, which is weighted by the perturbation probabilities $prob(p)$, given by

$$R_{a,P}^s = \int_{p \in P} prob(p) D_{a,P}^s d_p. \quad (7.2.3)$$

Here we propose to use the following measure to evaluate the average behavior

$$R_{a,P}^M = \sum_{i,j} \left[\int_{p \in P} prob(p) x_{ij}(p) d_p \right], \quad (7.2.4)$$

which is the mean $\overline{x_{ij}(p)}$ of gene expression levels over all perturbed model parameters. This means that it should be close to the simulated gene expression levels x_{ij} obtained from the unperturbed rate constants.

Depending on the evaluation function $D_{a,p}^s$, we can consider different measures to evaluate the robustness property. For example, if we consider the difference between the perturbed and unperturbed simulations, the impact of perturbations on dynamic behaviour can be defined by

$$R_{a,p}^M = \sum_{i,j} \left[\int_{p \in P} \text{prob}(p) (x_{ij} - x_{ij}(p))^2 d_p \right], \quad (7.2.5)$$

where $x_{ij}(p)$ and x_{ij} are the simulated gene expression levels at time point t_j with perturbed and unperturbed rate constants, respectively. For each set of model parameters, we generate M perturbed simulations and measure the difference between the perturbed and unperturbed simulations by

$$E_k = \sqrt{\frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M (x_{ij}^{(k)}(p) - x_{ij}^{(k)})^2} \quad (7.2.6)$$

where $x_{ij}^{(k)}(p)$ and $x_{ij}^{(k)}$ are perturbed and unperturbed k -th simulation of gene i at time point t_j , respectively.

In this work we are interested in the bistability property of the Nanog expression levels. The expression level may be in a low or high expression state. We need to consider the percentages of the system state staying in the low expression level, namely

$$P(\theta) = \int_{t \in \Omega_t} 1(|X(\theta) - X_L| \leq \epsilon) d_t \quad (7.2.7)$$

where $X(\theta)$ is the system state obtained using parameter θ and X_L is the low expression level of the system. For the different values of $P(\theta)$ based on given θ , we can use the mean and variance of these values to represent the robustness property of the system model.

$$R_M = E_\theta[P(\theta)], \quad R_V = \text{Var}_\theta[P(\theta)] \quad (7.2.8)$$

Here we formulate the following algorithm to calculate the sensitivity and robustness measure.

Algorithm for sensitivity and robustness analysis

Input: a stochastic model of chemical reactions with given parameters. Here calculate sensitivity measure (7.2.2) and robustness (7.2.7) related to parameter θ .

Step 1. Vary parameter θ to get a series of values $(\theta_1, \dots, \theta_n)$

Step 2. Use parameter θ_i to get M simulations of the stochastic model at time points T_1, \dots, T_K .

Step 3. Use these M simulation values to get a density function $f(x_k, \theta_i, T_j)$ at each time step T_j for each variable x_k .

Step 4. Apply the density function values $f(x_k, \theta_i, T_j)$ ($i = 1, \dots, n$) and a polynomial interpolation to calculate the derivatives of the density function $\frac{\partial f(x_k, \theta_i, T_j)}{\partial \theta_i}$.

Step 5. Obtain

$$\sum_{i=1}^n \sum_{k=1}^K \left| \frac{\partial f(x_k, \theta_i, T_j)}{\partial \theta_i} \right| f(x_k, \theta_i, T_j)$$

to get the sensitivity measure of the model against parameter θ_j .

Step 6. Repeat steps 1~5 to get the sensitivity measures for all parameters we are interested and then compare the influence of each parameter on the dynamics of the stochastic model.

Step 7. For robustness property, we use the generated M simulations in Step 2 to calculate either the difference between the perturbed and unperturbed simulations (7.2.6) or the percentages of the system state at a particular system state (7.2.7) as well as their mean and variance (7.2.8).

7.3 Results

7.3.1 Deterministic behaviour

In this work we first determined the dynamic behaviour of the deterministic model. Thus we derive the ordinary differential equation model based on the chemical reaction system (7.2.1), given by

$$\begin{aligned}
 \frac{d[\text{OS}]}{dt} &= \frac{s_{1,2}[\text{OS}]^2}{(1/k + [\text{OS}])^2 d_{o,s}} - d_{os}[\text{OS}] \\
 \frac{d[\text{N}]}{dt} &= \frac{s_3[\text{OS}]}{1/k + [\text{OS}]} + \frac{s_4[\text{N}]^2}{1/k + [\text{N}]^2 + p[\text{E}]} - d_N[\text{N}] \\
 \frac{d[\text{R}]}{dt} &= \frac{s_5[\text{OS}]}{1/k + [\text{OS}]} + \frac{s_6[\text{N}]^2}{1/k + [\text{N}]^2} - d_R[\text{R}] \\
 \frac{d[\text{E}]}{dt} &= \frac{s_7[\text{OS}]}{1/k + [\text{OS}]} - d_E[\text{E}]
 \end{aligned} \tag{7.3.1}$$

Similar to the stochastic model (7.2.1), variables in this deterministic model are molecule copy numbers. The rate constants are the same as those in Section 7.2.1 except parameters s_4 and p . As discussed in the next section, the values of these parameters are $s_4 = 150$ and $p = 250$. Note that model (7.3.1) is a special case of the model in (Herberg *et al.*, 2014) but it has two different parameter values.

For determining the function of intrinsic noise on the expression of gene Nanog, we investigate the existence of bistability regarding variations of parameters s_3 , s_4 , d_N and p . Since the enhanced degradation from Nanog is zero (namely $i_{OS} = i_N = i_R = 0$), the activity of Oct4-Sox2 is auto-regulated, and its steady state satisfies

$$\frac{s_{1,2}[\text{OS}]^2}{(1/k + [\text{OS}])^2 d_{o,s}} = d_1[\text{OS}]$$

which gives $[OS] = 7500$. Based on this steady state value, the steady state of FGF4/Erk complex is

$$[E] = \frac{1}{d_E} \frac{s_7[OS]}{1/k + [OS]} = 1.9973.$$

Thus, the steady states of Nanog gene will satisfy the following equation

$$\frac{7500s_3}{1/k + 7500} + \frac{s_4[N]^2}{1/k + [N]^2 + 1.9973p} = d_N[N] \quad (7.3.2)$$

Fig. 7.2 gives the bifurcation diagrams of the deterministic model (7.3.1). For all these four parameters, Nanog expression levels have two stable steady states and one unstable steady state over a range of parameter values. For parameters s_3 and p , the expression values at the steady states are relatively constant. The changes of these two parameters have not much influence on the expression level of Nanog gene unless p is very large. However, the synthesis rate s_4 and degradation rate d_N can change the steady states substantially. An interesting observation is that the unstable steady state is quite close to the steady state with low expression level. Thus, it is relatively easier for the system to switch from a low Nanog level to high Nanog expression level.

7.3.2 Stochastic behaviour

We use the proposed stochastic model to get simulations of Nanog gene network; and assume that the regulation strength p from PGF4/ERK complex is stochastic. Since the variables in (Herberg *et al.*, 2014) are molecular concentrations, we need to change some parameters to realize the variables in the model as molecular copy numbers in order to realize genetic switching for the expression level of gene Nanog. If we directly use parameters in Section 7.2.1, intrinsic noise in the stochastic model (7.2.1) is not large enough to induce Nanog to reach the other steady state. Thus we made some changes to some parameters, namely $s_4 = 150$

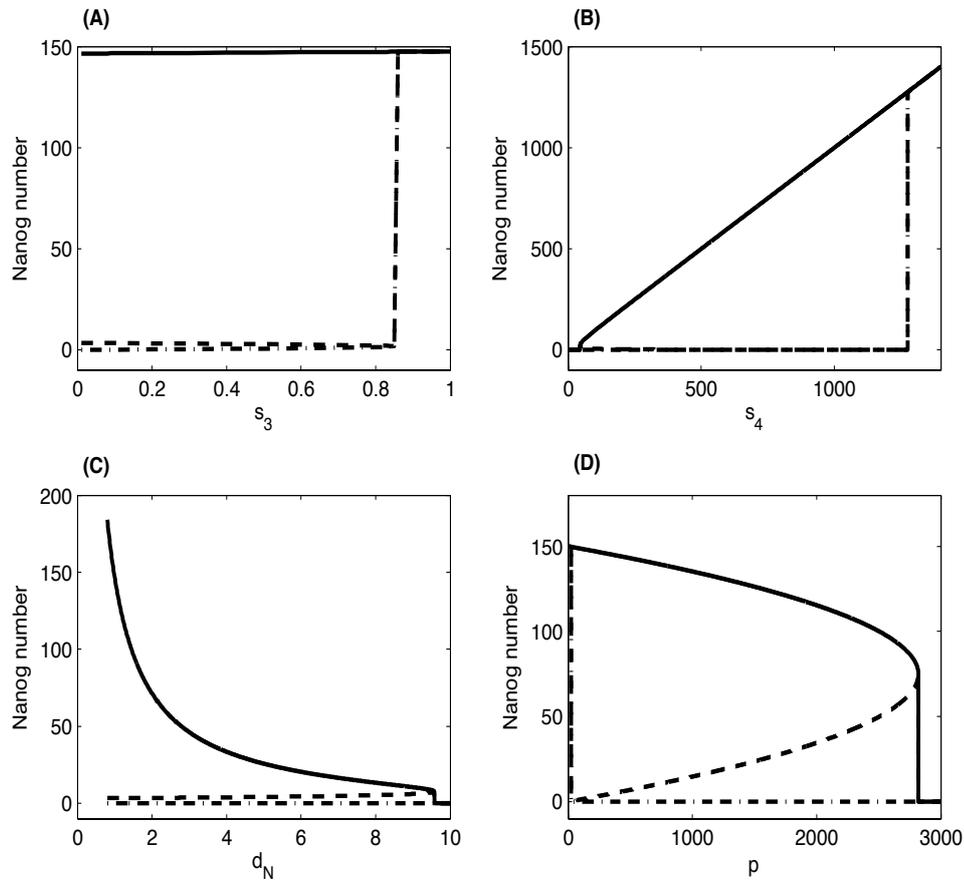


Figure 7.2: Bifurcation diagram of the deterministic model (7.3.1) for four parameters. Solid and dash-dot lines are two stable steady states but dash line in the middle is the unstable steady state. (A) parameter s_3 . (B) s_4 . (C) d_N . (D) p .

and

$$p = p_0 \cdot U(0,1) \quad (7.3.3)$$

where $p_0 = 500$ and $U(0,1)$ is a sample of the uniformly distributed random variable in $[0,1]$. Then we can realize genetic switching for the expression levels of Nanog and Rex1. Figure 7.3 gives one simulation of the system model. Fig. 7.3(A) presents the level of Oct4-Sox2 complex that always stay in a very high level. The Nanog expression level in Fig. 7.3(B) shows two steady states and noise can induce the system from one steady state to the other. Since gene Rex1 is regulated by Nanog, the expression levels of Nanog and Rex1 are at high or

low levels simultaneously, therefore expression level of Rex1 in Fig. 7.3(C) is consistent with that of Nanog in Fig. 7.3(B). However, the expression level of Rex1 is much less fluctuated than that of Nanog in Fig. 7.3(B). Finally the copy number of PGF4/ERK complex in Fig. 7.3(D) is quite low. The large variation in this complex number together with the fluctuations in the p value (7.3.3) can realize genetic switching.

Since parameter p is a random coefficient, the proposed model (7.2.1) is not purely based on intrinsic noise only. In fact, it is also influenced by external noise. We have also simulated the stochastic model using a fixed value of p , for example $p = p_0/2$ (results not shown). From Figure 7.3(D), we can see it is not easy to realize genetic switching using such constant value of parameter p , which is also suggested by the bifurcation diagram in Fig. 7.2(D).

7.3.3 Sensitivity analysis

In this section, we first calculate the density function of stochastic simulations obtained by different parameter values and then the corresponding derivatives of density functions. For each parameter in the model, we choose 11 different values of that parameter using

$$\theta_i = \theta_0 * (0.7 + i * 0.05), \quad i = 1, \dots, 11.$$

where θ_0 is the parameter value in Section 7.2.1. For each set of parameters, we obtain 1000 simulations and use the simulated expression levels at time point t to get the density distribution function.

Fig. 7.4 shows an example for the simulated density distributions of the four variables with different values of parameter d_E . Bell shapes of the density function are revealed for all the four species from this plot. Meanwhile the bell shaped

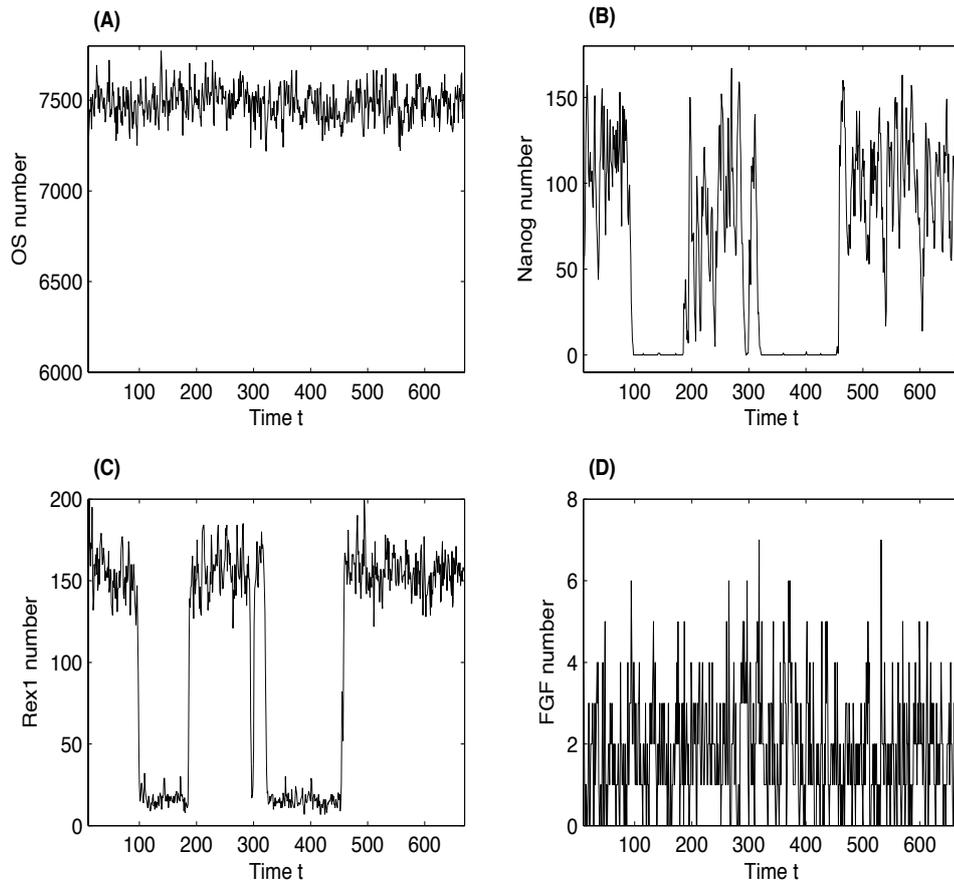


Figure 7.3: A stochastic simulation of the proposed stochastic model (7.2.1) using the first set of parameters together with $s_4 = 150$ and random variable p defined by (7.3.3).

density function can also be observed for other parameters (results not shown). In the simulated density functions in Fig. 7.4, the peak value of density function for different parameter values θ_i may be different, such as the E density function in Fig. 7.4 for the PGF4/ERK complex.

According to these density functions, the derivatives of each density function at different time points are also obtained. Since the partial derivative is defined for the parameter θ_i , we use a polynomial interpolation to approximate the density function in terms of the parameter and use the density function value at θ_{i-1} , θ_i

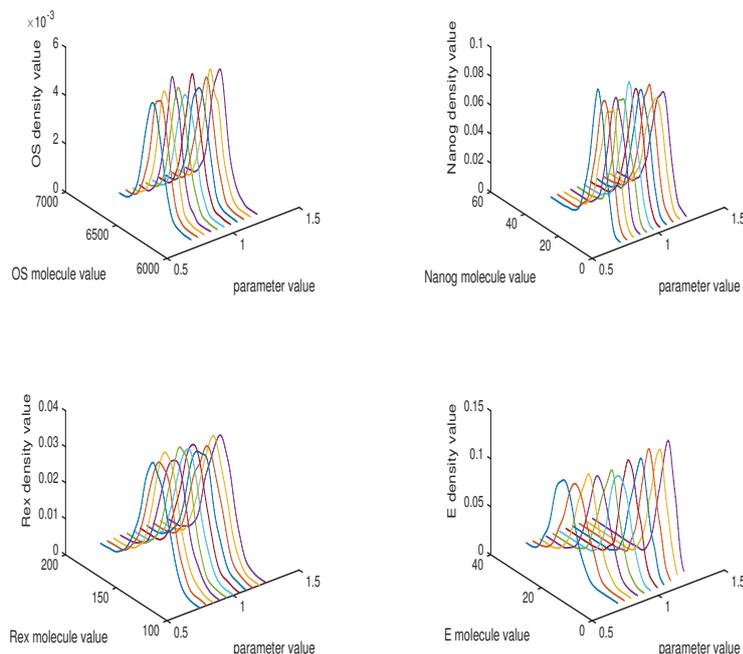


Figure 7.4: Density functions for variables [Oct4-Sox2, Nanog, Rex1, FGF4/Erk] with variations in parameter d_E

and θ_{i+1} to calculate the derivative of the density function at θ_i . Fig. 7.5 gives the derivative values of the density functions for parameter d_E .

Using the density function values in Fig. 7.4 and derivatives in Fig. 7.5, we then determine the sensitivity measure of four variables for different parameters at different time points. Fig. 7.6 shows that, for all the four variables, the variations of parameter s_3 has much larger influence on the system dynamics than all other parameters. However, the difference between the sensitivity measures of other parameters is small. This observation may be due to the large value of s_3 and large expression levels of Oct4-Sox2 complex.

To give a further indication of the sensitivity property, we first calculate the averaged values of the sensitivity measure over all the time points for each variable and each parameter. Then we sum up the averaged sensitivity values of the four

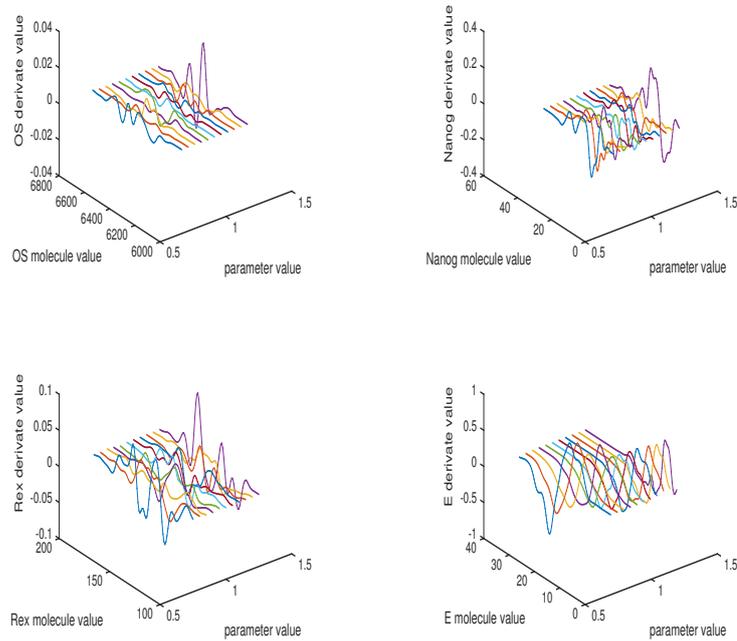


Figure 7.5: Derivatives of density functions for variables [Oct4-Sox2, Nanog, Rex1, FGF4/Erk] with parameter d_E

variables for each parameter. Figure 7.7 clearly shows that the variation of s_3 has much larger impact on the sensitivity property of the system than other parameters. In addition, the variation on any one of the four degradation rates has larger influence on the sensitivity property than the synthesis rate except s_3 . Surprisingly, the change of p_0 has slight influence on the sensitivity property.

7.3.4 Robustness analysis

After the assessment of sensitivity property of the stochastic model, we next investigate the effect of parameter changes on the system dynamics. In the Nanog gene network, we are interested in the bistability property of the Nanog expression level. Thus, we quantitatively measure the possibility for the Nanog gene to stay in a low expression level over a simulation. Compared with the expression levels of gene Nanog, the expression level of the marker gene Rex1 is consistent with the

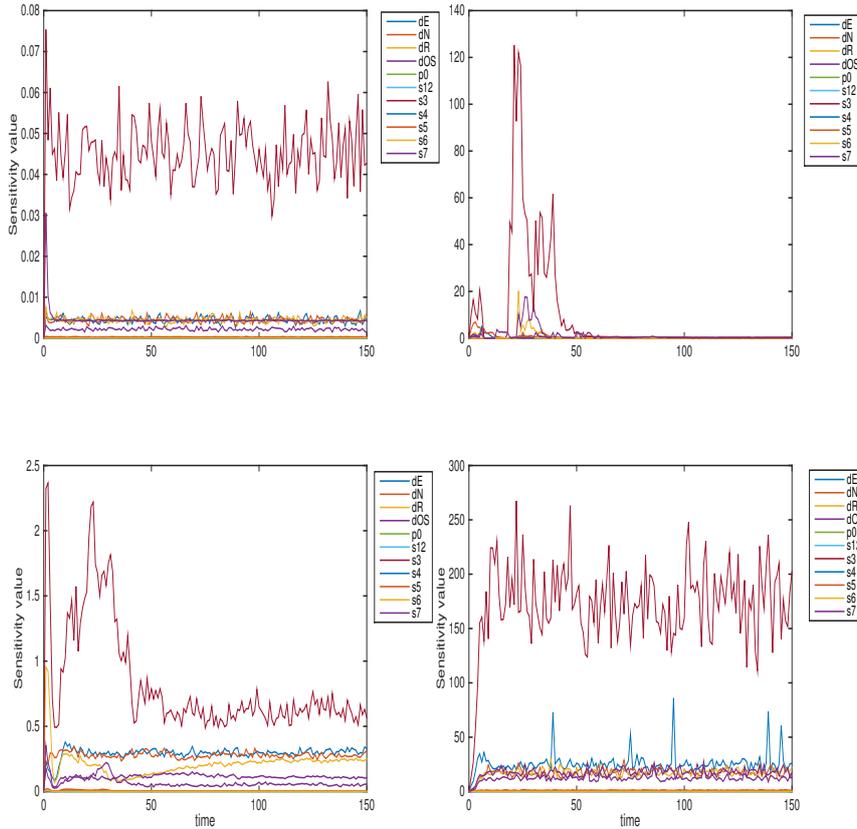


Figure 7.6: Sensitivity values for variables [Oct4-Sox2, Nanog, Rex1, FGF4/Erk] with all parameters

activity of Nanog but has less fluctuations. Thus we use the following formula to determine whether the expression level of Rex1 is in the low state at a given time point t .

$$[Rex1](t) < \min(Rex1) + \frac{1}{3}(\max(Rex1) - \min(Rex1)). \quad (7.3.4)$$

where $\min(Rex1)$ and $\max(Rex1)$ are minimal and maximal expression levels of Rex1 over a simulation.

For each set of model parameters, we obtained 1000 simulations and then calculated the percentages of gene Rex1 staying in the low expression levels at time point $t = 50$. Fig. 7.8 shows four types of changes for the influence of parameter variations on bistability properties. The first type of parameters, including s_{12} ,

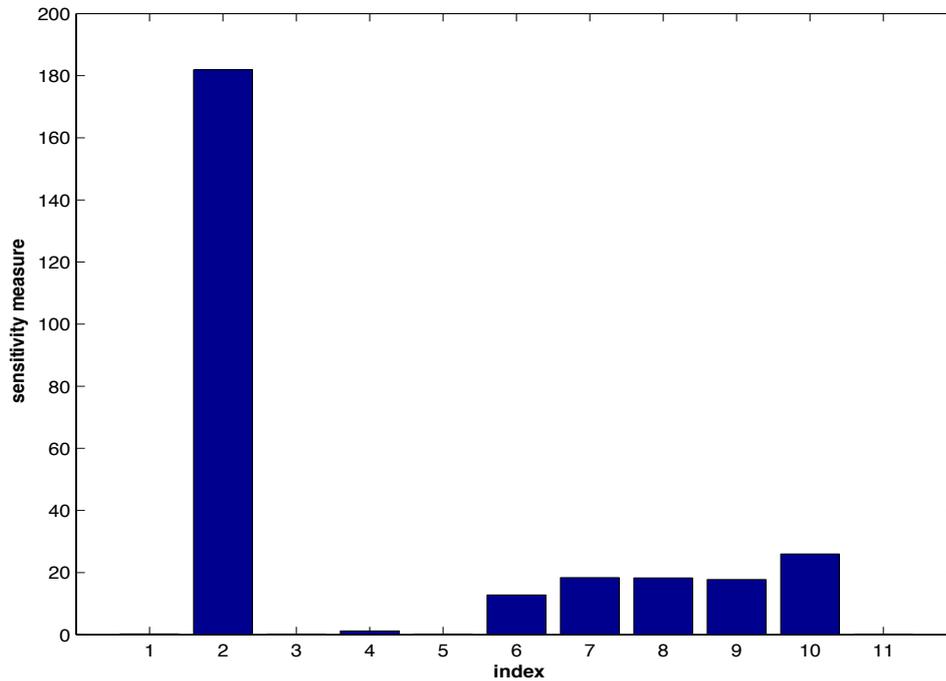


Figure 7.7: Sensitivity values of the stochastic model with all parameters. The indexes (1 ~ 11) are for (1: s_{12} , 2: s_3 , 3: s_4 , 4: s_5 , 5: s_6 , 6: s_7 , 7: d_{OS} , 8: d_N , 9: d_R , 10: d_E and 11: p_0)

s_5 , s_6 , d_{OS} , and d_R , does not have much influence on the bistability property of gene network. The percentages of Rex1 at low level always fluctuates around a particular value. However, an increasing of parameters d_N , p_0 and s_7 , which belongs to the second type, leads to more simulations whose expression levels of Rex1 stay at the low level at $t = 50$. On the contrary, increasing parameter value s_4 and d_E in the third type will promote the expression levels of Rex1, and more simulations in this case will maintain at the high expression level at $t = 50$. One special case can be noticed for parameter s_3 , i.e. instead of revealing a linear decreasing relationship, it presents some special patterns for the robustness of the system. In this case, which is the fourth type, the robustness property keeps unchanged in a small range of parameter variations. However, if the change of parameter value is large, the percentages of Rex1 staying at low level is also changed, though the variation of percentages is not high.

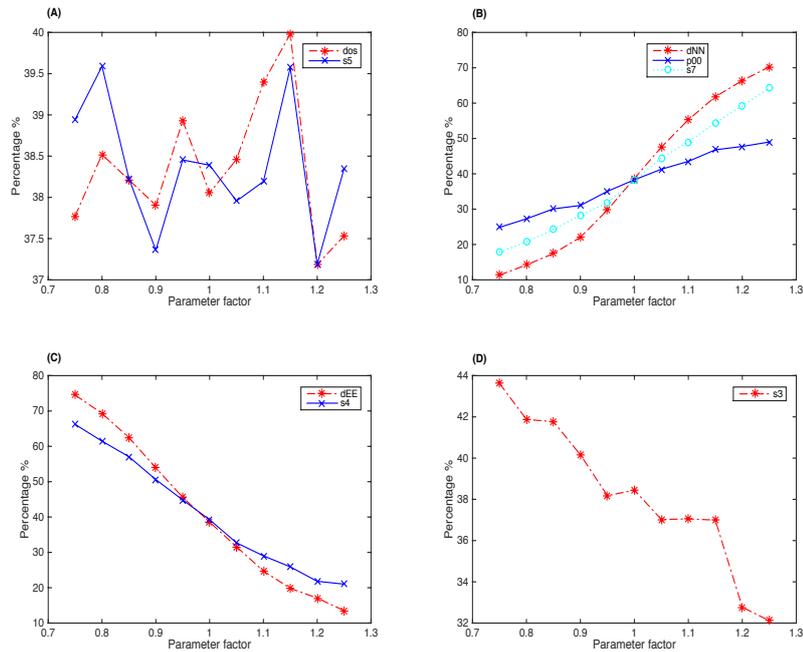


Figure 7.8: Robustness analysis showing the percentages of time points when the network maintains a low expression level of gene Rex1 using different values of a particular parameter. (A) Parameters d_{OS} and s_5 . (B) D_{NN} , p_{00} and s_7 . (C) d_{EE} and s_4 . (D) s_3

To further compare the sensitivity and robustness property, we give four simulations of Nanog with different values of s_3 or s_4 . Sensitivity analysis shows that the variation of parameter s_3 has a large impact on the system dynamics; while the influence of the changes in s_4 is small. Numerical simulations in Fig. 7.9 confirm this sensitivity analysis result. Fig. 7.9(A) and 7.9(B) give simulations when s_3 is 0.075 and 0.125, respectively. There are large variations in the Nanog number, though the percentage of Nanog in low expression level is relatively fixed according to the criterion (7.3.4). In fact it may be difficult to claim that the Nanog activity is in a low state since the time periods to maintain the high/low expression level are short. In addition, Fig. 7.9(C) and 7.9(D) give simulations when s_4 is 30 and 50, respectively. In this case, the Nanog copy number maintains in a high or low level for a relatively long time period. In addition, when the

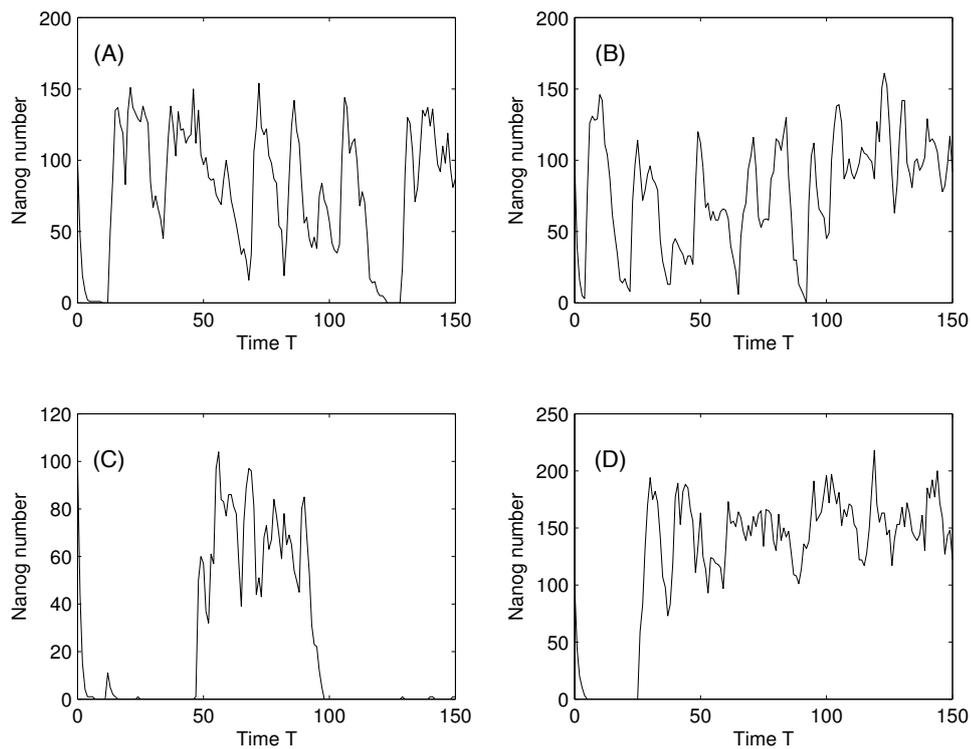


Figure 7.9: Simulations of Nanog number using different model parameters. (A) $s_3 = 0.075$. (B) $s_3 = 0.125$. (C) $s_4 = 30$. (D) $s_4 = 50$.

Nanog copy number stays in the high level, the variations of Nanog number are relatively smaller than those in Fig. 7.9(A) and 7.9(B). Thus both sensitivity and robustness property can provide a thorough understanding for the influence of model parameter variations on system dynamics.

7.4 Discussion and Conclusion

In this work, we proposed a framework to discuss the sensitivity and robustness property of biological systems simultaneously. In this framework sensitivity analysis uses the difference method to calculate the derivatives of the probability density function; and robustness analysis is based on the general definition proposed by Kitano (2007). Meanwhile a stochastic model of Nanog gene network with intrinsic noise based on a published model that discussed extrinsic noise only

was proposed. Numerical results of the Nanog gene network model suggest that the system dynamics are sensitive to variations of parameter s_3 that is related to the positive regulation from the Oct4-Sox2 complex. However, the system model is robust to the variation of this parameter. In addition, the change of a number of other parameters will vary the bistability property of the Nanog gene network model. Numerical simulations also indicated that the proposed framework is an efficient approach to assess the robustness and sensitivity properties of biological network models.

Sensitivity and robustness are two major concepts to measure the variation of system dynamics caused by parameter perturbations. Sensitivity measures quantitative changes of the variable values in the model; while robustness is the property of a system to maintain certain key properties, such as the bistability or oscillation. In the latter case, the quantitative variation of system output is not an important issue. Both are important to a biological system model under different experimental conditions. For the Nanog network model, our results raise a number of interesting questions regarding the sensitivity and robustness analysis, such as the quantitative definitions of these two properties and relationship between them. These issues will be the topics of potential future research.

PART B: Suggested Declaration for Thesis Chapter

Monash University

Declaration for Thesis Chapter 8

Declaration by candidate

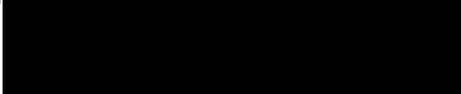
In the case of Chapter 8, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Verified the method, wrote part of the programming codes and the article	30%

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Junbai Wang	Developed, established and verified the method Wrote programming codes and the article	
Tianhai Tian	Provided helpful guidance and editorial work	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature		Date 29/04/15
Main Supervisor's Signature		Date 29/04/15

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Chapter 8

An Integrated Approach to Infer Dynamic Protein-gene Interactions: A Case Study of the Human P53 Protein

Chapter 8 is based on the article Wang J, Wu Q, Tian T. 2015. An integrated approach to infer dynamic protein-gene interactions: a case study of the human p53 protein (submitted for publication) .

Abstract. *Investigating the dynamics of regulatory networks through high throughput experimental data, such as microarray gene expression profiles, is a very important but challenging task. One of the major hindrances in building detailed mathematical models for genetic regulation is the large number of unknown model parameters. To tackle this challenge, a new integrated method is proposed by combining a top-down approach and a bottom-up approach. First, the top-down approach using probability graphical models is used to predict the network structure of DNA repair pathway that is regulated by the p53 protein. Two networks are predicted, namely a network of eight genes with eight inferred interactions and that of 21 genes with 17 interactions. Then, the bottom-up approach using differential equation models are developed to study the detailed genetic regulations based on either a fully connected regulatory network or a gene network obtained by the top-down approach. Model simulation error, Akaike's information criterion, parameter identifiability and robustness property are used as criteria to select the optimal network. Results based on random permutation of input gene network structures indicate that the prediction accuracy and robustness property of the two inferred networks are much better than that of the corresponding fully connected networks. In addition, a comparison study suggests that the proposed approach has better simulation accuracy and robustness property than the earlier one. In particular, the computational cost is significantly reduced. Overall, the new integrated method is a promising approach for investigating the dynamics of genetic regulation.*

Keywords. *P53; gene regulatory network.*

References are considered at the end of the thesis.

Chapter 8

An Integrated Approach to Infer Dynamic Protein-gene Interactions: A Case Study of the Human P53 Protein

8.1 Introduction

With the rapid advancement of high-throughout technologies such as microarray and mass spectrometry (MS)-based proteomics, it has become possible to measure gene expression levels and kinase activities in the genome-wide scale simultaneously (Wang, 2008; Cox and Mann, 2011; Simon, 2008). Although the datasets contain enormous amounts information of biological systems, it is still a challenge to develop effective methods to extract useful knowledge from the observations (Rung and Brazma, 2013). In particular, inference of genetic regulatory networks is considered as an important task for extracting hidden information from observations such as microarray gene expression datasets. The development of inference

methodologies and their application to genetic networks have attracted substantial interests from researchers in a wide range of research fields and become one of the important topics in bioinformatics (Bar-Joseph *et al.*, 2012; Penfold and Wild, 2011; Zhang *et al.*, 2013; Wang *et al.*, 2011).

Mathematical modelling and statistical analysis have contributed substantially to the basic understanding of biological processes. Inference methods, which identify molecular interaction networks from ‘omics’ datasets, are termed as top-down approaches in systems biology (Bruggeman and Westerhoff, 2007). Data mining and machine learning techniques have been used to infer interactions or correlations among various variables. Inference methods for gene regulatory networks based on time-course gene expression profiles include Pearson correlation (Nayak *et al.*, 2009), the Boolean networks (Hickman and Hodgman, 2009), Gaussian graphical models (Wang *et al.*, 2005; Ma *et al.*, 2007), Bayesian networks (Wang and Li, 2012; Friedman *et al.*, 2000), Bayesian correlated clustering (Kirk *et al.*, 2012), models based on support vector machines (Zhu *et al.*, 2009), and singular value decomposition (Yeung *et al.*, 2002). Usually, probabilistic graphical model (i.e. Gaussian graphical models or Bayesian networks) is the first choice because of its simplicity and efficiency (Maetschke *et al.*, 2013).

Another major type of methods for inferring gene networks is the bottom-up approach that examines the mechanisms through functional properties of the interactions between network components (Bruggeman and Westerhoff, 2007). Among the bottom-up approaches, differential equation model that is based on the available information of variables (i.e. TF) is particularly important. This approach not only captures the dynamic behavior of gene expression but also provides more detailed regulatory information than the top-down approaches (Gardner *et al.*, 2003). Since the bottom-up method has many unknown parameters that need to be estimated using experimental data, it only suites for small-scale networks. Generally, a linear model is the first choice to infer a large gene network

because of the simplicity and computing efficiency (De Jong, 2002). However, the linear model is not appropriate for studying systems with non-linear properties (Tegner *et al.*, 2003). Although an S-system can be used to realize non-linear properties of genetic regulations (Savageau, 1969; Thomas *et al.*, 2007), it has a large number of unknown model parameters. More recently, stochastic differential equations have been used to describe the function of noise in microarray gene expression data (Tian, 2010; Wang and Tian, 2010), and new algorithms have been developed to infer the network structure and parameters of mathematical models for the bottom-up approaches (Liu and Wang, 2008; Vilela *et al.*, 2008; Cao and Zhao, 2008; Akutsu *et al.*, 2000; Kimura *et al.*, 2005; De Smet and Marchal, 2010).

For bottom-up approaches such as differential equations models, inference is generally defined as a problem of estimating model parameters that produce small simulations errors against experimental data. Nevertheless, a real hindrance of genetic network inference is that the number of unknown model parameters is much larger than that of observation time points using gene expression profiles. For example, in the linear model for a network of N genes, $N(N + 1)$ unknown parameters have to be estimated based on observations within at most 50 time points. In an S-system model, the number of unknown parameters is $2N(N + 1)$; and the neural network model holds $N(N + 3)$ unknown parameters. Therefore, a key challenge for the bottom-up approaches is to reduce the number of unknown model parameters from the level of N^2 to the level of N . So far, several approaches have been proposed to address this issue, including the usage of network properties (i.e. network structural sparseness, network scale-freeness, as well as network motif and modularity) (Hecker *et al.*, 2009), but the results were still not satisfactory.

Although either top-down or bottom-up approaches have been used separately to infer gene regulatory networks in various system scales, there has been little integrative research so far to combine both methods together. In fact, the inferred gene-gene interactions (either positive or negative) through top-down

approaches may assist the fine tuning of unknown parameters in a bottom-up method. Thus, we propose a new method by combining both top-down and bottom-up approaches together to infer genetic regulatory networks. The new method may not only reduce the dimension of parameter space but also save computational time.

8.2 Materials and Methods

8.2.1 Experimental dataset

This research is based on a published microarray dataset that was generated from the Human All Origin MOLT4 cells carrying wild-type p53. Cell were irradiated and harvested every 2 hours over a 12-hours period (Barenco *et al.*, 2006). We obtained the ionizing radiation Affymetrix dataset from ArrayExpress (E-MEXP-549). Pre-processing of microarray datasets were based on a previous publication (Wang and Tian, 2010) such as probes with bad signal quality and less variation across all the time points were removed. A pair-wise Fisher's linear discriminant method (Wang *et al.*, 2003a) was used to screen probes with the most relevant response to ionizing radiation. Based on a previous developed modelling method, a total of 317 putative target genes of the p53 proteins was predicted (Wang and Tian, 2010). From the predicted target genes, genes related to the DNA repair system were selected for this research. Among them, eight most relevant response genes were selected from the top 100 putative target genes in (Wang and Tian, 2010) to form a small-scale gene network; and 21 most relevant response genes from the top 317 potential target genes were selected to form a medium-scale gene network. More information of these putative target genes can be found in the Supplementary Information in (Wang and Tian, 2010).

The dataset provides microarray gene expression profiles at seven time points. To test the model quality as well as to ensure for reliable inference results, we use a spline function to approximate gene expression levels at every other time points. A similar approach has been used to generate more time point measurements based on the raw microarray expression dataset (Bar-Joseph *et al.*, 2003). In every two hours time interval, we add three time points with equal distance that resulted in 25 time points with time-step-size 0.5. Visualization of gene networks is realized by Cytospace software.

8.2.2 Top-down approach

Probabilistic graphical model

The probabilistic graphical model is a probability model for multivariate random observations whose independence structure is characterized by an independence graph (Wang *et al.*, 2003b). For example, a graph G is a mathematical object that consists of two sets, a set of vertices K , and a set of edges E that consists of pairs of elements taken from K . If all edges are undirected then the graph is undirected. The undirected independence graph gives a picture of the pattern of dependence or association between the variables.

Gaussian graphical model with a forward search algorithm (GGF)

Given an independence graph G and a k -dimensional continuous random vector X with a multivariate normal distribution, a covariance selection model (Wang *et al.*, 2005) was used to search for the best independence graph. In the GGF algorithm, the conditional independence constraints are equivalent to specifying zeros in the parameters in the inverse of the covariance matrix corresponding to the absence of an edge in G . In other words, two variables are independent given the remaining variables if and only if the corresponding element of the inverse of the covariance

matrix is zero. A more detailed description of the computing process for the GGF algorithm is presented in Figure 8.5. Here we give a brief description of the GGF that is used in the current work as follows:

- 1. Let $X = (X_1, X_2, \dots, X_k)$ be a k -dimensional vector, k be the number of genes. An initial empty graph G is built where k vertices correspond to k genes.
- 2. An iterative maximum likelihood estimates algorithm (Dempster *et al.*, 1977) is used to compute the covariance matrix, $\text{Cov}(G)$, of the initial graph G .
- 3. An edge E_i is added into the initial graph and then a new covariance matrix, denoted as $\text{Cov}(E_i)$, is estimated by the iterative maximum likelihood estimates. The significance of the added edge is tested by the deviance difference, which has an asymptotic Chi-square distribution with one degree of freedom. A P -value of the Chi-square test is used as the model selection criteria.
- 4. If the P -value of an added edge E_i is smaller than a predefined cut-off P -value (e.g. significance level $P < 0.05$), the edge is added to the initial graph G , and then go back to step 2. It is reiterated between step 2 and step 4 until the P -value of added edge is larger than a threshold value.
- 5. Based on the inferred undirected graph from step 4, graph orientation rules are applied to transform it into a directed acyclic graph (DAG).
- 6. In the final DAG, vertices represent genes; edges depict the association between a pair of genes; arrows explain possible causes and effects between a pair of genes; and the correlation coefficient between a pair of vertices tells positive or negative association between two genes.

The GGF can be used to predict gene-gene interactions when the number of time points is much less than the number of genes.

8.2.3 Bottom-up approach

Mathematical model

We have proposed a general framework to describe the dynamics of gene expression (Wang and Tian, 2010). For a network with N genes and M transcriptional factors (TFs), we denote the expression levels of i -th gene as $x_i(t)$. The dynamics of gene expression is represented by the following differential equations, given by

$$\frac{dx_i}{dt} = c_i + k_i f_i(x_1(t - \tau_{i1}), \dots, x_N(t - \tau_{iN}), P_1(t - \tau_{i(N+1)}), \dots, P_M(t - \tau_{i(N+M)})) - d_i x_i, \quad (8.2.1)$$

for $i = 1, \dots, N$, where c_i and k_i are the basal transcriptional rate and maximal expression rate of gene i , respectively, d_i is transcript degradation rate of gene i , $[P]_j$ is the activity of the j -th TF, and τ_{ij} is regulatory delay of gene j related to the expression of gene i . The regulatory function $f_i(x_1, \dots, x_N, P_1, \dots, P_M)$ includes both positive and negative regulations. In this work we propose to use the following function

$$f_i = \frac{\sum_{j=1}^N a_{ij} x_j^{n_{ij}}(t - \tau_{ij}) + \sum_{j=1}^M a_{i(N+j)} [P]_j^{n_{i(N+j)}}(t - \tau_{i(N+j)})}{1 + \sum_{j=1}^N b_{ij} x_j^{n_{ij}}(t - \tau_{ij}) + \sum_{j=1}^M b_{i(N+j)} [P]_j^{n_{i(N+j)}}(t - \tau_{i(N+j)})} \quad (8.2.2)$$

Coefficients a_{ij} ($j = 1, \dots, N + M$) represent regulations from gene j ($1 \leq j \leq N$) or TF j ($N + 1 \leq j \leq N + M$) to the expression of gene i . This regulation may be positive ($a_{ij} > 0$) or negative ($a_{ij} = 0$) if the corresponding coefficient ($b_{ij} > 0$). There will be no regulatory relationship from gene j to gene i if ($a_{ij} = b_{ij} = 0$). Note that these assumptions imply that ($a_{ij} \neq 0$) only when ($b_{ij} > 0$). When ($a_{ij} > 0$), it is assumed that gene i can autoregulate the expression of itself.

In the network studied in this work, it contains N genes and only one TF ($M = 1$), namely the p53 proteins that regulate all these genes either positively or negatively. It has been established that the p53 proteins form tetramers as TFs. Thus the exponent 4 in (8.2.3) represents the tetramer structure of p53 as TFs, namely $n_{i(N+1)} = 4$. When ($a_{i(N+1)} > 0$), p53 regulates the expression of gene i positively. However, if ($a_{i(N+1)} = 0$), the expression of gene i is inhibited by p53.

In addition, for simplicity we do not consider time delay in the expression levels of genes (i.e. $\tau_{ij} = 0, i, j = 1, \dots, N$) and cooperative binding (i.e. $n_{ij} = 1, i, j = 1, \dots, N$). Thus the synthetic function used in this paper is

$$f_i = \frac{a_{i1}x_1 + \dots + a_{iN}x_N + a_{i(n+1)}[P(t - \tau_i)]^4}{1 + b_{i1}x_1 + \dots + b_{iN}x_N + b_{i(n+1)}[P(t - \tau_i)]^4} \quad (8.2.3)$$

and $k_i = 1$. Here $[P]$ is the activity of TF p53 whose detailed dynamics are presented in Fig. 8.1 in (Wang and Tian, 2010), and τ_i is the regulatory delay of TF p53 related to the expression of gene i . The time delay is estimated by our proposed algorithm and its value is available in the Supplementary Information (Wang and Tian, 2010).

Inference methods

All model parameters are estimated using the genetic algorithm, which is an effective searching method for finding the unknown kinetic rates when the search space is associated with a complex error landscape. We used a MATLAB toolbox (Chipperfield *et al.*, 1994) to infer the unknown model parameters. This toolbox used MATLAB functions to build a set of versatile routines for implementing a wide range of genetic algorithms. The major procedures using genetic algorithm toolbox include population representation and initiation, fitness assignment, selection functions, crossover operators, mutation operators and multiple sub-population support. In this work we used the function *crtbp* to create the

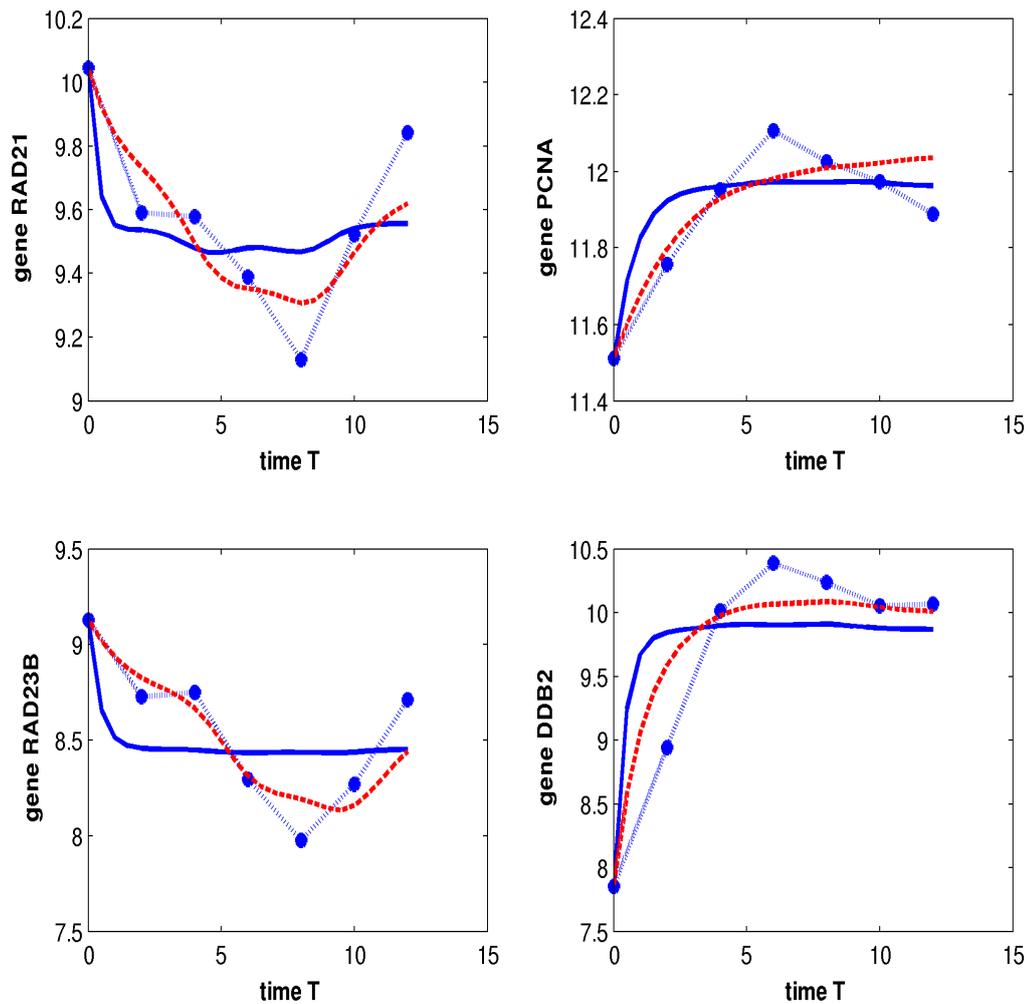


Figure 8.1: Simulations of the gene network of eight genes: *RAD21*, *pcnA*, *RAD23B*, *DDB2* (dash-star: microarray data; solid-line: simulation of the fully connected model; dash-line: simulation of the core network predicted by the GGF).

binary initial population, the linear-ranking and non-linear-ranking algorithms to transform the raw objective function values into non-negative figures of merit for each individual, a selection function *reins* to effect fitness-based reinsertion when the entire population is not reproduced in each generation, a high-level entry function *select* to provide a convenient interface to the selection routines, a high-level entry function *recombine* to provide all the crossover operators, and the routine *mut* to perform binary and integer mutations.

The genetic algorithm is ran over 500 or 1000 generations for each estimate of model parameters, and we use a population of 100 individuals in each generation. The values of rate constants are taken initially from the uniform distribution in the range of $[0, W_{max}]$, and the value of W_{max} for parameters $(a_{ij}, b_{ij}, c_i, k_i, d_i)$ are $(20, 20, 20, 20, 1)$, respectively. Here the value of W_{max} is determined by numerical tests. For each parameter, we first select an initial value of W_{max} to infer model parameters. If certain estimates are very close to W_{max} , the value of W_{max} is increased; however, the value of W_{max} is decreased if the estimated values are substantially smaller than W_{max} . The initial estimate of rate constants can be changed using different random seeds in MATLAB, leading to different final estimates of rate constants. For each mathematical model, we infer 20 sets of model parameters and select the top five sets with minimal errors for further analysis.

Model measurement criteria

The error of an inferred set of model parameters is measured by the residual sum of squares between the simulated expression levels and experimental data, defined by

$$E = \sum_{i=1}^N \sum_{j=1}^M (x_i(t_j) - x_{ij})^2, \quad (8.2.4)$$

where x_{ij} and $x_i(t_j)$ are the simulated and experimental expression levels of gene i ($i = 1, \dots, N$) at time point t_j ($j = 1, \dots, M$), respectively.

This work considers a number of mathematical models with different regulatory mechanisms that are realized by different model coefficients in model (8.2.1). As the model complexity or the number of model parameters increases, the model becomes more capable of adapting to the characteristics of expression data. To address the issue of over-fitting, we use Akaike's Information Criterion (AIC) to measure the quality of mathematical model. AIC is the first model selection criterion, which was designed to extend the maximum likelihood principle (Akaike, 1974). The traditional maximum likelihood paradigm provides an approach for estimating unknown parameters of a model having a specific dimension and structure. AIC extends this paradigm by considering a framework in which the model dimension is also unknown. For small sample size, namely when $L/d < 40$, AIC is defined by the following equation (Symonds and Moussalli, 2011)

$$AIC = L \log(V) + 2d + \frac{2d(d+1)}{L-d-1}, \quad (8.2.5)$$

where d is the number of estimated parameters, L the number of values in estimation dataset, and V the loss function that is defined by the residual sum of squares

$$V = \frac{1}{L} \sum_{i=1}^N \sum_{j=1}^M (x_i(t_j) - x_{ij})^2, \quad (8.2.6)$$

where x_{ij} and $x_i(t_j)$ are simulated and experimental measured data of gene i at time point t_j , respectively.

In addition, we examine parameter identifiability of the model based on the inferred model parameters. Parameter identifiability is a property of a model structure that ensures that parameters can be uniquely (globally or locally) determined from knowledge of the input-output behaviour of the system. Here a

model is not identifiable if different sets of parameter values result in the same model. More specifically, let $M(\theta)$ be the function that defines a model, which has unknown parameters θ , and then a model is globally identifiable if

$$M(\theta_1) = M(\theta_2) \quad (8.2.7)$$

implies that $\theta_1 = \theta_2$. In addition, a model is locally identifiable if there exists an open neighbourhood of any θ such that the identifiability (8.2.7) is true. Otherwise a model is non-identifiable (Cole *et al.*, 2010). In recent years, a number of approaches have been proposed to analyse parameter identifiability of biological system models (Raue *et al.*, 2014; Hines *et al.*, 2014).

For an inferred model parameter set θ_0 , we consider the local identifiability of the model in the neighbourhood of parameter set θ_0 . Let $h(x|\theta)$ be the model prediction that ensures that the residual sum of square

$$S = \sum_{l=1}^n [y_l - h(x_l|\theta)]^2 \quad (8.2.8)$$

has a unique minimum. Here $\{(x_i, y_i)\}_{i=1}^n$ are specified dataset at n points and θ is a parameter vector with dimension p . We consider a matrix H whose element is defined by

$$H_{ij} = \frac{\partial h(x_i|\theta)}{\partial \theta_j} \quad (8.2.9)$$

$$i = 1, \dots, n, j = 1, \dots, p.$$

Then the model is local identifiable in the neighbourhood of parameter set θ_0 when the matrix $H^T H$ has full rank ($= p$) (Little *et al.*, 2010). Since there is no overlap between parameters in different equations in model (8.2.1), we examine the identifiability property for each equation separately. We further assume that the prediction $h(x|\theta)$ is made using the implicit Euler method for solving model

(8.2.1). Thus the element H_{ij} is the partial derivative of the right-hand side of the i -th differential equation in model (8.2.1) with respect to each parameter θ_j .

Robustness analysis

Robustness, in both biological and engineering systems, is defined as the ability of a system to function correctly in the presence of both internal and external uncertainty (Csete and Doyle, 2002). Since robustness is a ubiquitously observed property of biological systems (Kitano, 2004; Tian and Song, 2012), this property has been widely used recently as an important measure to select the optimal network structure or model rate constants from estimated candidates (Citri and Yarden, 2006; Apri *et al.*, 2010; Masel and Siegal, 2009). A formal and abstract definition of the robustness property (Kitano, 2007) is used in this work to measure the robustness property of the proposed model. The robustness property of a mathematical model with respect to a set of perturbations P is defined as the average of an evaluation function $D_{a,P}^s$ of the system over all perturbations $p \in P$, which is weighted by the perturbation probabilities $prob(p)$, given by

$$R_{a,P}^s = \int_{p \in P} prob(p) D_{a,P}^s dp. \quad (8.2.10)$$

Here we propose to use the following measure to evaluate the average behavior

$$R_{a,P}^M = \sum_{i,j} \left[\int_{p \in P} prob(p) x_{ij}(p) dp \right], \quad (8.2.11)$$

which is the mean $\overline{x_{ij}(p)}$ of gene expression levels over all the perturbed model parameters. This means that it should be close to the simulated gene expression levels x_{ij} obtained from the unperturbed rate constants. In addition, the impact

of perturbations on dynamic behaviour is defined by

$$R_{a,P}^M = \sum_{i,j} [\int_{p \in P} \text{prob}(p)(x_{ij} - x_{ij}(p))^2 dp], \quad (8.2.12)$$

where $x_{ij}(p)$ and x_{ij} are the simulated gene expression levels at time point t_j with perturbed and unperturbed rate constants, respectively.

For each rate constant k_i , the perturbation is set to

$$\bar{k}_i = \max\{k_i(1 + \mu N), 0\} \quad (8.2.13)$$

with the standard Gaussian random variable $N(0, 1)$. Here μ represents the perturbation strength and we have tested various values of μ . Numerical results suggest that if the value is small, noise does not have much impact on the system dynamics and the drift of simulations is small for all models. So in this work we use $\mu = 0.2$ in the robustness analysis.

For each module of gene regulation, we use the genetic algorithm to generate 20 sets of model parameters, and then select the top 5 sets that have the minimal estimation error for robustness analysis. In this way, we are able to exclude the influence of simulation error on the robustness property of the model. For each set of model parameters, we generate 1000 perturbed simulations and measure the difference between the perturbed and unperturbed simulations by

$$E_k = \sqrt{\frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M (x_{ij}^{(k)}(p) - x_{ij}^{(k)})^2} \quad (8.2.14)$$

where $x_{ij}^{(k)}(p)$ and $x_{ij}^{(k)}$ are perturbed and unperturbed k -th simulation of gene i at time point t_j , respectively. Robustness and STD in Tables 8.1 ~ 8.5 are the mean and standard deviation of E_k for each network model.

8.3 Results

8.3.1 Inference of a network of eight genes with full connections

To demonstrate the importance of top-down approach in the inference of genetic regulation, we first examine a common approach by assuming each genes in the network may regulate the expression of all other genes. In this fully connected regulatory network, matrices $A = [a_{ij}]_{N \times N}$ and $B = [b_{ij}]_{N \times N}$ in model (8.2.3) are full matrices; and the number of unknown model parameters in the network model of N genes is $2N^2 + 5N$. This approach has been applied to infer small-scale or medium-scale gene networks (De Smet and Marchal, 2010). To form a small-scale gene network, we first select eight genes from the top 100 predicted putative p53 targeted genes, namely *RAD21*, *pcnA*, *RAD23B*, *DDB2*, *PTTG1*, *XPC*, *RAD51C*, and *Rps27L* that are all related to the DNA repair pathway. Genetic algorithm was used to search for optimal model parameters that minimize simulation errors to microarray gene expression data. For this network, the number of unknown model parameters is 168. The ratio of a_{ij}/b_{ij} determines the relative influence of the activity of gene j on the expression of gene i . We also assume that the value of a_{ij} is zero if this ratio is below a threshold value, and in this case gene j negatively regulates the expression of gene i . If the values of both a_{ij} and b_{ij} are under a threshold, then gene j has no influence on the expression of gene i . The inferred model parameters are listed in the Supplementary Information Table (8.6).

Figure (8.1) shows the simulated expression profiles of four genes in the network, and the remaining four genes are given in the Supplementary Information Fig. (8.6). Here, the simulated expression levels match the experimental data very well. The ratio of a_{ij}/b_{ij} is below the assumed threshold value for only a few

regulations, which suggests that the majority of predicted regulations are positive. However, from the gene expression profiles, the regulations between some pairs of genes may be negative. In addition, simulation results show that both a_{ij} and b_{ij} are below the threshold value only in limited cases. Table (8.1) shows that for this fully connected network, all robust measurements (i.e. AIC and robustness property) are larger than those of networks with less model parameters in the following sections.

8.3.2 Inference of the network of eight genes using predicted network structure from top-down approach

To improve the robustness property of gene network and reduce the computational time, GGF is used to predict network structure of gene-gene interactions. Since the predicted network structure depends on the significant level of GGF, the algorithm is applied multiple times on the same eight genes with different p -values. For example, from p -value 0.009 to p -value 0.05, the inferred network structure remains the same, which is also true when the p -value is increased from 0.09 to 0.2. Thus, two predicted gene networks by GGF are finally chosen for subsequent analysis using bottom-up approaches, one with $p < 0.05$ (Fig. (8.2A) with 8 mutual regulations) and the other with $p < 0.09$ (Fig. (8.2B) with 17 mutual regulations).

Based on the predicted network structure in Fig. (8.2A) with 8 mutual regulations, genetic algorithm is used to search for optimal model parameters. The model has only ~ 72 unknown parameters, which is much less than that of the fully interacted network model. However, the predicted model still has adequate flexibility to realize the experimental data (i.e. 56 and 200 measurements in the raw and extended data, respectively). The value of a_{ij} ($i, j = 1, \dots, N$) is determined by the graphic model, namely its value is either positive or zero if the regulation is

positive or negative. The inferred model parameters are given in Supplementary Information Table (8.7). Although the graphic model has much less potential regulations and consequently less unknown model parameters, simulations in Fig. (8.1) and Fig. (8.5) in Supplementary Information suggest that the error of the graphic model is slightly smaller than that of the fully connected network. This result indicates that the addition of more model parameters does not reduce the estimation errors. It is worth to note that less accuracy of the fully connected model does not mean this model has less capacity to match the characteristics of expression data. In fact, as the searching space is more complicated when more model parameters are added, it may be more difficult to search for optimal model parameters.

The AIC and robustness property of the graphic model are also tested. Table (8.1) shows that the graphic model has much better property of AIC than the fully connected network model. However, the fully connected model has slightly better robustness property than the graphic model. In addition, our results suggest that for the inferred model parameter sets, the network model with 8 mutual regulations is parameter identifiable when the extended dataset with 25 measurement time points are used. However, it is not parameter identifiable when the raw microarray dataset with 7 measurement time points are used.

8.3.3 Inference of the network of eight interactions with extended regulations

We have successfully used the graphic model with only eight mutual interactions ($P < 0.05$) to realize the experimental data with good accuracy and system property. The next question is whether the network of eight genes in Fig. (8.2A) is the core network. To answer this question, we examine the possibility of adding mutual regulations derived from an extended network ($P < 0.09$) in Fig. (8.2B)

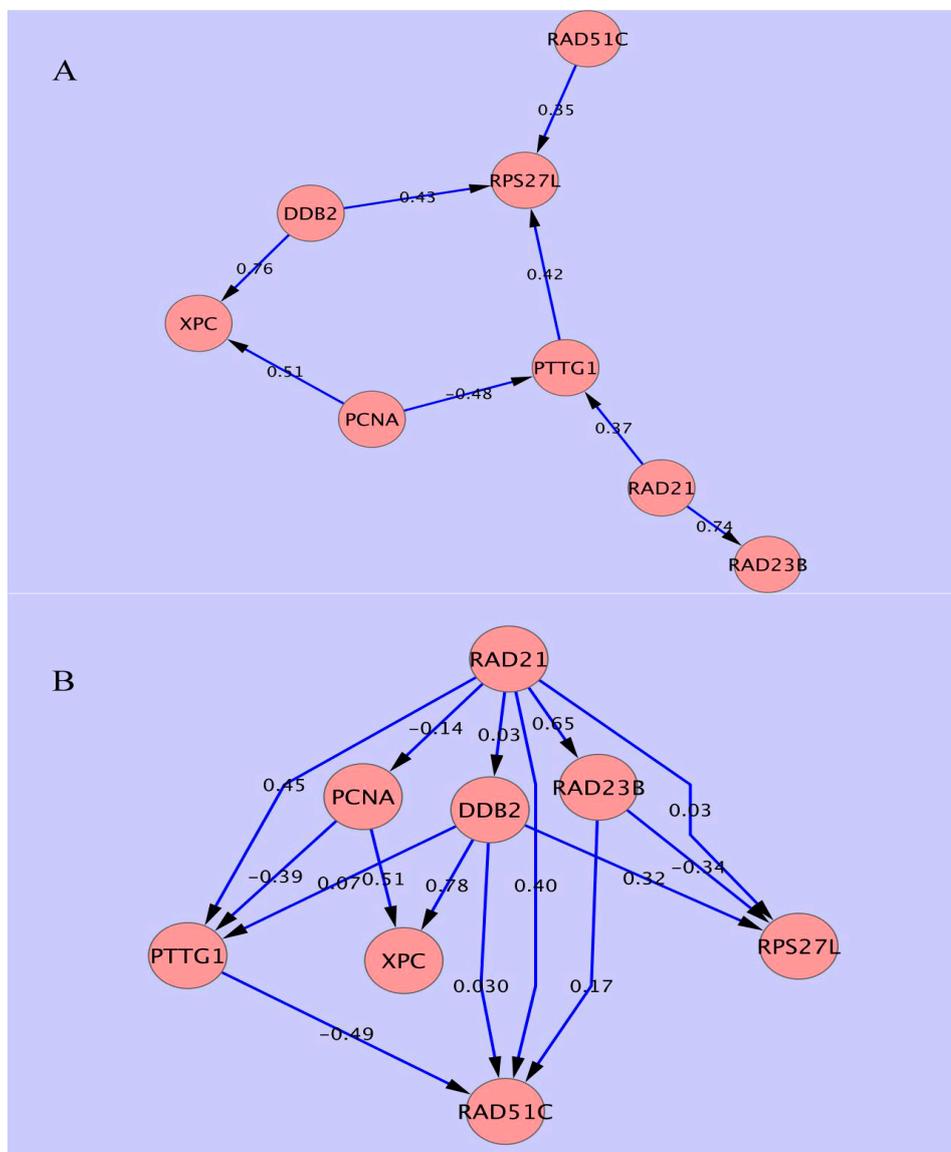


Figure 8.2: The graphic models of eight genes using different significant levels: Gene-gene interaction networks are predicted by applying GGF on 8 genes that are related to DNA repair pathway: (A) there are 8 regulations among 8 genes, significance level $p < 0.05$; (B) there are 17 regulations among 8 genes, significance level $p < 0.09$, in which includes 8 regulations from the (A). Number on each edge is partial correlation coefficient between the two genes. The network is visualized by Cytoscape software.

Table 8.1: AIC and robustness property of mathematical model of eight genes: Numerical results are presented as the AIC to experimental data, average behaviour of robustness property and standard deviation (STD) of robustness property, which are based on the average of five sets of estimated parameters. (G_i, G_j) means the network by adding the mutual regulation of gene i and gene j to the core network ($p < 0.05$). "Plus 5 regulations" is the network by adding 5 edges, namely $(G4, G5)$, $(G1, G4)$, $(G3, G7)$, $(G5, G7)$, $(G1, G2)$. ($G1$: RAD21, $G2$: *pcnA*, $G3$: RAD23B, $G4$: DDB2, $G5$: PTTG1, $G6$: XPC, $G7$: RAD51C, $G8$: *Rps271*)

Network	AIC	Robustness	STD
Fully connected network	-61.69	2.723	1.162
Core network of 8 genes ($p < 0.5$)	-112.65	3.110	1.234
(G1, G8)	-180.08	3.138	1.721
(G4, G5)	-110.24	3.137	1.234
(G1, G4)	-111.30	3.092	1.307
(G3, G7)	-109.46	3.053	1.259
(G4, G7)	-107.02	3.139	1.464
(G1, G7)	-108.53	3.181	1.730
(G3, G8)	-109.89	3.099	1.277
(G5, G7)	-113.10	3.161	1.283
(G1, G2)	-108.28	3.113	1.370
plus 5 regulations	-99.16	3.029	1.085
Extended model ($p < 0.09$)	-96.96	2.984	1.205

to the network ($P < 0.05$) in Fig. (8.2A). Since there are nine additional mutual regulations in the network with ($P < 0.09$), nine extended models are considered by adding one mutual regulation in each test. Genetic algorithm is used to infer model parameters in the extended models in order to realize experimental data. For each extended model, the top 5 sets with smaller simulation errors of all 50 estimates are selected for robustness analysis. AIC values and robustness analysis results are shown in Table (8.1), which suggest that the AIC values of the extended models are close to that of the core model in Fig. (8.2A). In addition, more model parameters in the extended models contribute to larger AIC values. The parameter identifiability analysis suggests that each extended model is parameter identifiable when the extended dataset with 25 measurement points are used. However, these extended models are not parameter identifiable when the raw microarray dataset with 7 measurement points are used, which is the same as the core model. Regarding the robustness property, Table (8.1) suggests that, if one edge is added, only three extended models have better robustness property than the core model while the standard deviation of robustness property for all extended models are not as good as the core model; however, the extended two models with more edges are better than the core model in terms of robustness property.

Next, we select five mutual regulations that have either smaller simulation errors (even though the AIC value of the model with extended regulation is larger than the core model) or better robustness property. We then test the effect of adding all these five regulations into the core network. Results in Table (8.1) suggest that the network with five added regulations has better robustness property than the core network, though the AIC value of this extended network is larger than that of the core model. We also test the extended network ($P < 0.09$) in Fig. (8.2B) with 17 regulations. Table (8.1) suggests the network in Fig. (8.2B) has larger AIC value but slightly better robustness property than the core model.

8.3.4 Inference of the network of eight regulations with auto-regulation

In both undirected and directed graphic model, the auto-regulation, namely the positive or negative regulation of a gene to the expression of itself, is not considered. To find the potential auto-regulation, we test the network by adding either positive or negative auto-regulation to each gene. For each gene i , we set $b_{ii} > 0$; and the value of a_{ii} is $a_{ii} > 0$ and $a_{ii} = 0$ for positive and negative auto-regulation, respectively. Similarly, we obtain 50 sets of estimated model parameters for each network with only one auto-regulation. The addition of one auto-regulation does not change the identifiability of the model, namely each extended model is parameter identifiable when the generated dataset with 25 measurement points are used; while they are not parameter identifiable when the raw microarray dataset with 7 measurement points are used. Table (8.2) suggests that the addition of four negative auto-regulations does decrease AIC values. However, among these four auto negative regulations, robustness analysis results show that only one negative auto-regulation, namely G7(-) in Table (8.2), increases the robustness property of the network. In addition, the addition of positive auto-regulation increases AIC values but decreases robustness property of the core model. Thus, numerical results suggest that a few negative auto-regulations may be added to the core network model. Nevertheless, the conclusion is derived from a special network in which TF *p53* protein regulates the expression of all genes in the network. A different conclusion may be derived when a different network model is considered.

Table 8.2: AIC and robustness property of the network ($p < 0.05$) by adding an auto-positive/negative regulation to the core network: $G_i (+)$: add auto-positive regulation of gene i to the core network; $(-)$: add auto-negative regulation. RBN: robustness.

	AIC	RBN	STD		AIC	RBN	STD
G1 (+)	-109.12	3.105	1.805	G1 (-)	-111.62	3.048	1.270
G2 (+)	-110.97	408.00	1.27E4	G2 (-)	-113.19	3.179	1.687
G3 (+)	-112.72	3.042	1.261	G3 (-)	-110.28	3.119	2.002
G4 (+)	-111.16	3.035	1.156	G4 (-)	-113.48	3.108	1.553
G5 (+)	-107.18	3.248	1.835	G5 (-)	-111.91	3.142	1.943
G6 (+)	-110.72	3.279	1.941	G6 (-)	-114.29	3.209	1.395
G7 (+)	-110.45	3.114	1.253	G7 (-)	-113.18	3.031	1.100
G8 (+)	-108.56	3.167	1.749	G8 (-)	-109.71	3.462	7.997

8.3.5 Network structure perturbation - edge deletion

After testing the addition of potential regulations to the core network ($P < 0.05$), the possibility of removing certain interactions from the network in Fig. (8.2A) is also examined. In the core network with eight genes and only 8 regulations, on average each gene has only two connections to the other genes. Particularly, some genes such as gene *RAD51C* have only one connection. Thus, only two types of regulation deletion are tested. First, we consider eight networks where one of the mutual regulations is removed from the core network. Using a similar method for parameter inference and robustness analysis, Table (8.3) suggests that removal of any mutual regulations can neither improve simulation accuracy nor decrease AIC values. Although edge (PTTG1, Rps27I) may be removed from the network due to decreased AIC value, robustness property of the reduced network is not as

good as the core network. Thus, no edge deletion is recommended because there is no convincing reason to remove any edge from the core regulatory network.

In the proposed model, if there is a potential regulation between gene i and gene j , it is assumed that both $a_{ij}(b_{ij})$ and $a_{ji}(b_{ji})$ are positive. This assumption is valid if the relationship between gene i and j are connective. However, the regulation between the two genes may be one-way regulation if it is transcription regulation. In the second test of regulation deletion, we test 16 cases of reduced model based on the core network with 8 regulations to test for potential one-way regulation. In each reduced model, we remove one of the one-way regulations of a mutual regulation. Table (8.4) shows that the AIC values of the reduced models are smaller than that of the core network model due to a smaller number of model parameters and smaller simulation errors of some reduced models. However, the removal of one one-way regulation also reduces the robustness property of the network model. There is no reduced model whose robustness properties are better than that of the core network model. These results suggest that there is a strong possibility for the one-way regulation only between certain pairs of genes in the network.

A widely used approach in network inference is to remove a potential regulation by checking the value of corresponding coefficient in the network model. Another important question remains now is whether the removed regulations identified by our studies can be selected by the inferred fully connected network model with small model parameters, namely a_{ij} and b_{ij} . To answer the question, we check the potential reduced models identified in Tables (8.3 and 8.4). Results show that there is not any consistence between the inferred removable regulation in these tables and small values of coefficient a_{ij} and b_{ij} in the inferred fully connected network model. However, we should mention that the methods for regulation deletion and those for model inference are quite different.

Table 8.3: AIC and robustness property of the reduced network with 8 genes by removing one mutual regulation from the core model in Fig. (8.2A): (G_i, G_j) means the network by removing mutual regulation between gene i and gene j from the core network.

	AIC	Robustness	STD
(G1, G3)	-116.11	3.151	1.601
(G1, G5)	-117.95	3.124	2.359
(G2, G5)	-115.30	3.135	1.464
(G2, G6)	-120.23	3.142	1.566
(G4, G6)	-119.81	3.200	2.054
(G4, G8)	-116.27	3.050	1.608
(G5, G8)	-116.26	3.102	1.251
(G7, G8)	-117.69	3.047	1.666

8.3.6 A Comparison study to an earlier inference method

To demonstrate the effectiveness of our proposed method, we conduct a comparison study using an ODE-based approach for inferring genetic regulatory networks from time-series and/or steady-state measurements (Äijö and Lähdesmäki, 2009). This approach is based on the use of Bayesian analysis with ODEs and non-parametric Gaussian process modelling for the transcriptional-level regulation. We input the expression data of the eight genes into the software in (Äijö and Lähdesmäki, 2009). The output is an 8×8 matrix whose diagonal elements are zeros. The (i, j) element represents the posterior probability of which gene j is regulated by gene i . All the derived elements are positive, thus it is assumed that all regulatory relationships between these genes are positive. To compare with the network in Fig. (8.2A) with 8 mutual regulations, we select top 16 elements that

Table 8.4: AIC and robustness property of the reduced network with 8 genes by removing one one-way regulation from the core model in Fig. (8.1A): ($G_i \leftarrow G_j$) represent the removing of the regulation from gene j to gene i , namely by letting $a_{ij} = b_{ij} = 0$. RBN: robustness.

	AIC	RBN	STD		AIC	RBN	STD
$G1 \leftarrow G3$	-118.30	3.319	3.416	$G1 \rightarrow G3$	-113.74	3.159	2.132
$G1 \leftarrow G5$	-112.91	3.264	1.986	$G1 \rightarrow G5$	-115.60	3.331	4.559
$G2 \leftarrow G5$	-117.51	3.125	1.488	$G2 \rightarrow G5$	-113.13	3.312	2.514
$G2 \leftarrow G6$	-116.73	3.300	4.085	$G2 \rightarrow G6$	-116.33	3.107	1.515
$G4 \leftarrow G6$	-116.54	3.614	9.424	$G4 \rightarrow G6$	-116.70	3.176	1.602
$G4 \leftarrow G8$	-116.32	3.150	1.636	$G4 \rightarrow G8$	-116.88	3.301	1.835
$G5 \leftarrow G8$	-117.06	3.202	1.819	$G5 \rightarrow G8$	-115.85	3.199	1.566
$G7 \leftarrow G8$	-116.93	3.134	1.131	$G7 \rightarrow G8$	-117.06	3.202	2.483

have the largest values of posterior probabilities in the matrix since the inferred network is directional. These 16 elements represent 3 mutual regulations and 10 one-way regulations, and the connection network is plotted in Fig. (8.8). The network in Fig. (8.8) shares two mutual regulations and three one-way regulations with that in Fig. (8.2A). Then we use the proposed model (8.2.1) and (8.2.3) to infer rate constants for the network in Fig. (8.8) and calculate AIC values. Table (8.5) shows that the inferred network model from Fig. (8.8) generate simulations that have larger AIC value than the core model in Fig. (8.2A). We also carry out robustness analysis for this network model. Table (8.5) suggests that our network model in Fig. (8.2A) is more stable than that in Fig. (8.8).

The eight genes considered in this work can be classified as two groups according to their expression patterns. The first group contains genes *RAD21*, *RAD23B* and *PTTG1*, which are negatively regulated by *p53*; while the remaining five genes are positively regulated by *p53*. In Fig. (8.2A), there are two regulatory relationships

Table 8.5: Simulation error and robustness property of mathematical models of eight genes: (Network 1: the gene network inferred from the network structure in Fig. (8.2A). Network 2: the gene network inferred from the network structure in Fig. (8.7); Network 3: the gene network inferred from the network structure that is merged from the networks in Fig. (8.2A) and Fig. (8.7)).

Network	AIC	Robustness	STD
Network 1	-112.65	3.176	1.668
Network 2	-102.60	3.384	1.218
Network 3	-86.32	3.375	1.180

between these two groups, namely (*pcnA*, *PTTG1*) and (*PTTG1*, *Rps27L*). The relationship between *pcnA* and *PTTG1* has been identified as a negative regulation using GGF algorithm, though the regulation between *PTTG1* and *Rps27L* is positive. In the network in Fig. (8.8), there are also other four one-way relationships between these two gene groups. However, all these four relationships are identified as positive regulations by a published method (Äijö and Lähdesmäki, 2009). Fig. (8.9) suggests that this network is not capable to produce simulations that are close to the expression levels of genes that are negatively regulated by *p53*.

To examine potential redundancy of regulation relationship between the two inferred networks, we merge these two networks together to form a network with nine mutual regulations and eight one-way regulations. Then we infer a network model for the emerged network. Simulations in Fig. (8.10) suggest that this network model does not realize expression levels of genes *RAD21* and *RAD23B* accurately. In addition, Table (8.5) suggests that the accuracy and robustness property of this merged network model are not as good as those of the network model in Fig. (8.8).

8.3.7 Inference of a medium network of 21 genes

With the success of a small network of eight genes, we further conduct an inference work for a network with 21 genes. In addition to the 8 genes in the previous network, the remaining 13 genes are *TOP2A*, *tp53*, *CIB1*, *papd7*, *GADD45A*, *FANCA*, *RBM14*, *H2afx*, *lig3*, *Mutyh*, *REV3L*, *Recql4*, and *IGHMBP2*. All the selected genes are related to DNA repair pathway and have strong response to ionizing radiation, though some of them may not have the same predicted quality as the eight genes (Wang and Tian, 2010). GGF algorithm is applied several times on the 21 genes using different significant p -values. Generally, there is a small variation of predicted gene network structure when $p < 0.009$. Note that gene *tp53* is not connected to the inferred network when $p < 0.009$. However, when $p > 0.009$, gene *tp53* is linked to all the rest genes and a fully connected network is obtained. This result indicates the importance of gene *tp53* in DNA repair pathway. Nevertheless, for simplicity, we select the network in Fig. (8.3) with $p < 0.009$ in the subsequent analysis.

Two potential networks have been studied for the network of 21 genes. The first one is the fully connected regulatory network. The number of unknown model parameters is $2N^2 + 5N = 987$, but the available data have only 147 and 525 measurements based on the raw microarray data and extended dataset, respectively. We obtain 50 sets of estimated model parameters and find that some model parameters have a wide range of estimated values. After examining the values of a_{ij} and b_{ij} as well as the ratio a_{ij}/b_{ij} , we find it is difficult to remove any regulation from the system since some parameters may have small values in one set of estimate but large values in another.

The second network in Fig. (8.3) is predicted by the GGF. It has only 26 pairs of regulations and 188 unknown model parameters. Genetic algorithm is used

to infer the optimal model parameters. Fig. (8.4) presents simulations for the expression levels of four genes; and Fig. (8.7) also gives simulations for the other four genes. It shows that the simulation of the GGF predicted network has provided more accurate simulations than the fully connected network. The robustness analysis of these two network models suggests that the GGF predicted model with less regulations and less model parameters shows better robustness property than that of the fully connected model.

8.4 Discussions

In this work we proposed a novel approach for inferring genetic regulatory network by combining both top-down and bottom-up approaches. To address issues regarding multiple regulations and sparseness of network topology, we first used the probabilistic graphic models to infer network structure. By choosing various significant levels of the graphic models, we derived a core network ($p < 0.05$) and an extended network ($p < 0.09$). The latter is a possible expansion of the core network. To validate the predicted graphic models and also to investigate detailed dynamic regulations, we designed a new mathematical model to represent complex regulation with both positive and negative regulations, regulations of TFs, protein cooperative binding, and time delay. Using our designed mathematical model, we first tested a fully connected regulatory network. Compared with the fully connected network, our predicted core network has smaller AIC values, parameter identifiability property and better robustness property. Subsequently, we tested the possibility of adding regulations to or removing regulations from the core network model, and used AIC values and robustness property as key criteria to validate the predicted models. We also conducted a comparison study using a published inference method. Numerical results suggested that our proposed

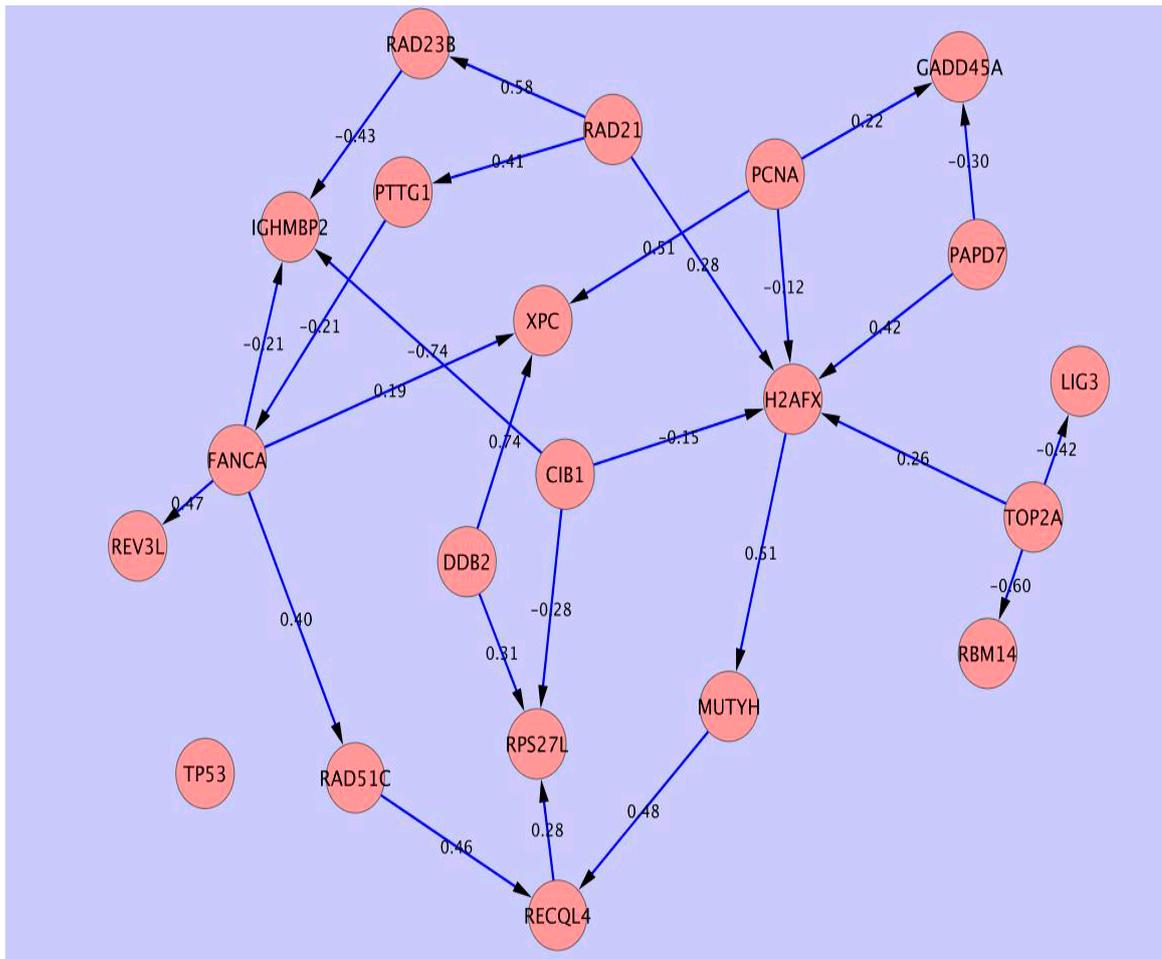


Figure 8.3: A gene-gene interaction network was predicted by applying GGF on 21 genes that related to DNA repair pathway: There are 26 regulations among the genes, significance level $p < 0.009$. Number on each edge is partial correlation coefficient between the two genes. The network is visualized by Cytoscape software.

method could predict network models that have better simulation accuracy and robustness property.

The underlying methods behind top-down and bottom-up approaches are different. For example, the GGF algorithm, which is a top-down approach, uses the correlation analysis method to find the regulatory relationship between pairs of genes; while the ODE-model (bottom-up approach) considers the expression of a gene that is regulated by a number of other genes. Thus inferred networks from these two types of methods are different. Currently, bottom-up approaches ignore the network structure derived from the top-down approaches. Instead

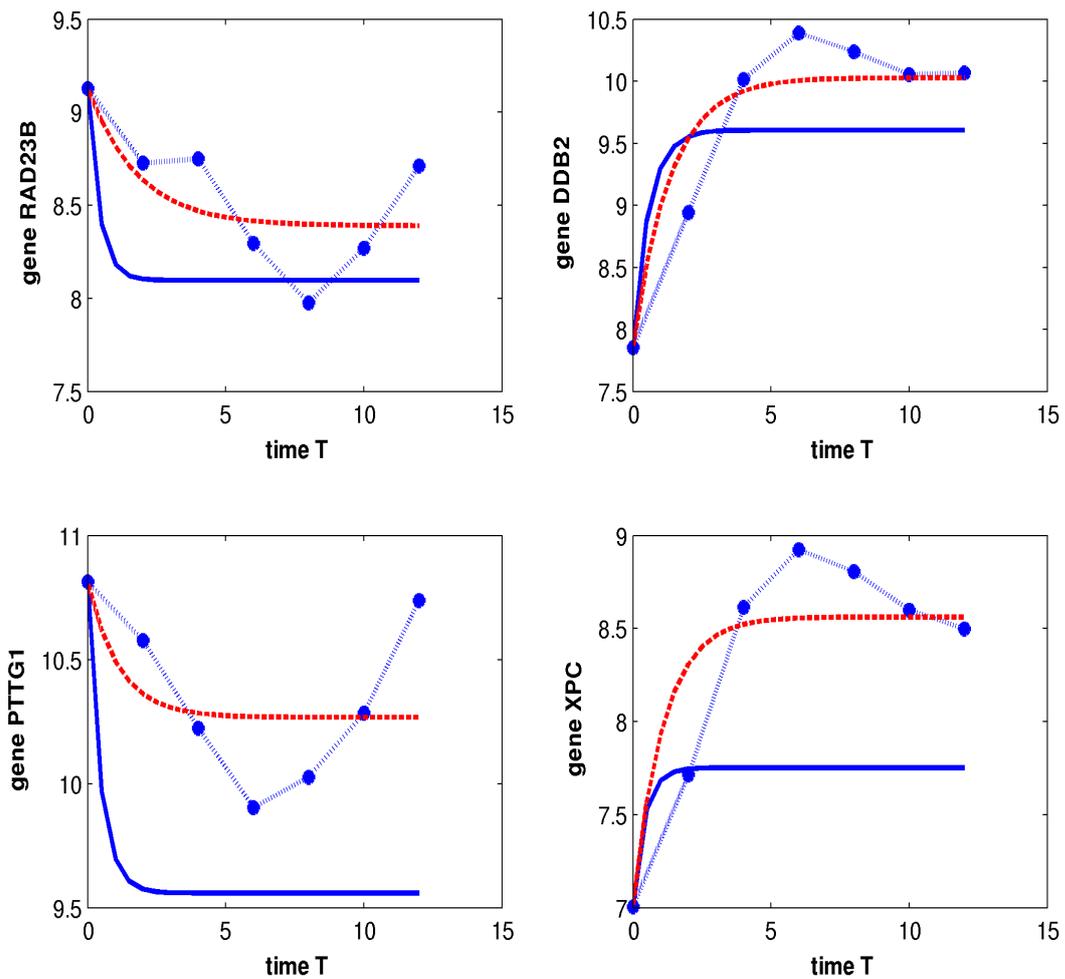


Figure 8.4: Simulations of the gene network with 21 genes (dash-star: microarray data; solid-line: simulation of the fully connected model; dash- line: simulation of the core network predicted by the GGF).

researchers infer gene networks from scratch using unsupervised, supervised or semi-supervised inference methods based on experimentally established networks. Thus an important question is to explore the consistency between the networks derived from the two types of inference methods. To address this issue, we developed a novel framework to infer regulatory networks using both types of inference methods. Our research results suggest that top-down approaches can provide initial simplified network structure, and more importantly, improve the accuracy and efficiency of the bottom-up approaches.

This research work raised a number of major issues regarding the implementation of these methods. One issue is that a fully connected regulatory network actually leads to simulations with larger errors compared with models with less unknown model parameters. Theoretically a model with more unknown parameters should provide more flexibility to fit observation data than a model with less unknown parameter. However, this does not mean a model is better if it has more unknown parameters. In fact, due to the limited amount of experimental data, a mathematical model with an adequate number of parameters may be already capable to realize experimental data with good accuracy. The addition of more parameters will not increase model flexibility much but will increase model complexity. For example, when using an optimization method to infer unknown parameters, more model parameters will increase the difficulty to search the optimal parameters because of the issue of local maximum of the optimization methods. In addition, when we say a model with more unknown parameters may be more flexible to fit experimental data, this statement is based on the assumption that the inference method can find parameters with zero value. However, the majority of current approaches fail to infer parameters with zero value. The estimated unknown parameters normally are non-zero. Thus a model with more unknown parameters may not be able to produce more accurate simulation than that with less model parameters, and it may be difficult to compare a model with more parameters with

that having less model parameters directly. Furthermore, because of the complex searching space of model parameters and noise in experimental data, it may be difficult to judge which model is better if the difference between their simulation errors is small. For example, simulation errors of various models for the network of eight genes are quite close to each other. Therefore, in addition to using simulation error as the unique criterion to select a model, other measurements, such as ACI value, parameter identifiability and robustness property of a network, are also needed as important criteria.

Due to issue of local maximum in optimization methods, we may obtain a number of estimates with varying parameter values. All these estimates can faithfully realize experimental data (Tian *et al.*, 2007b). An alternative approach is the Bayesian inference method that does not only estimate confidence intervals, but also provide even more information by estimating the whole posterior parameter distribution. However, one obstacle with the standard Bayesian approaches is the difficulty of exploring the huge discrete state space with a complicated likelihood structure that makes conditional simulation difficult. Recently, interest has been increasingly turned to methods that avoid some of the problem complexity by using forward-simulation methods such as likelihood-free Markov chain Monte-Carlo and approximate Bayesian computation (ABC). A potential next step is to develop effective Bayesian algorithms for inferring regulatory networks using large-scale omics datasets.

8.5 Supplementary Information

This following section includes Figure (8.5) for a work flow chart of Gaussian Graphical Model with Forward search algorithm (GGF). Figure (8.6) gives simulations of the remaining 4 genes in the core network model of eight genes ($p < 0.05$). Figure (8.7) presents the simulation of the nine genes in the network of 21 genes.

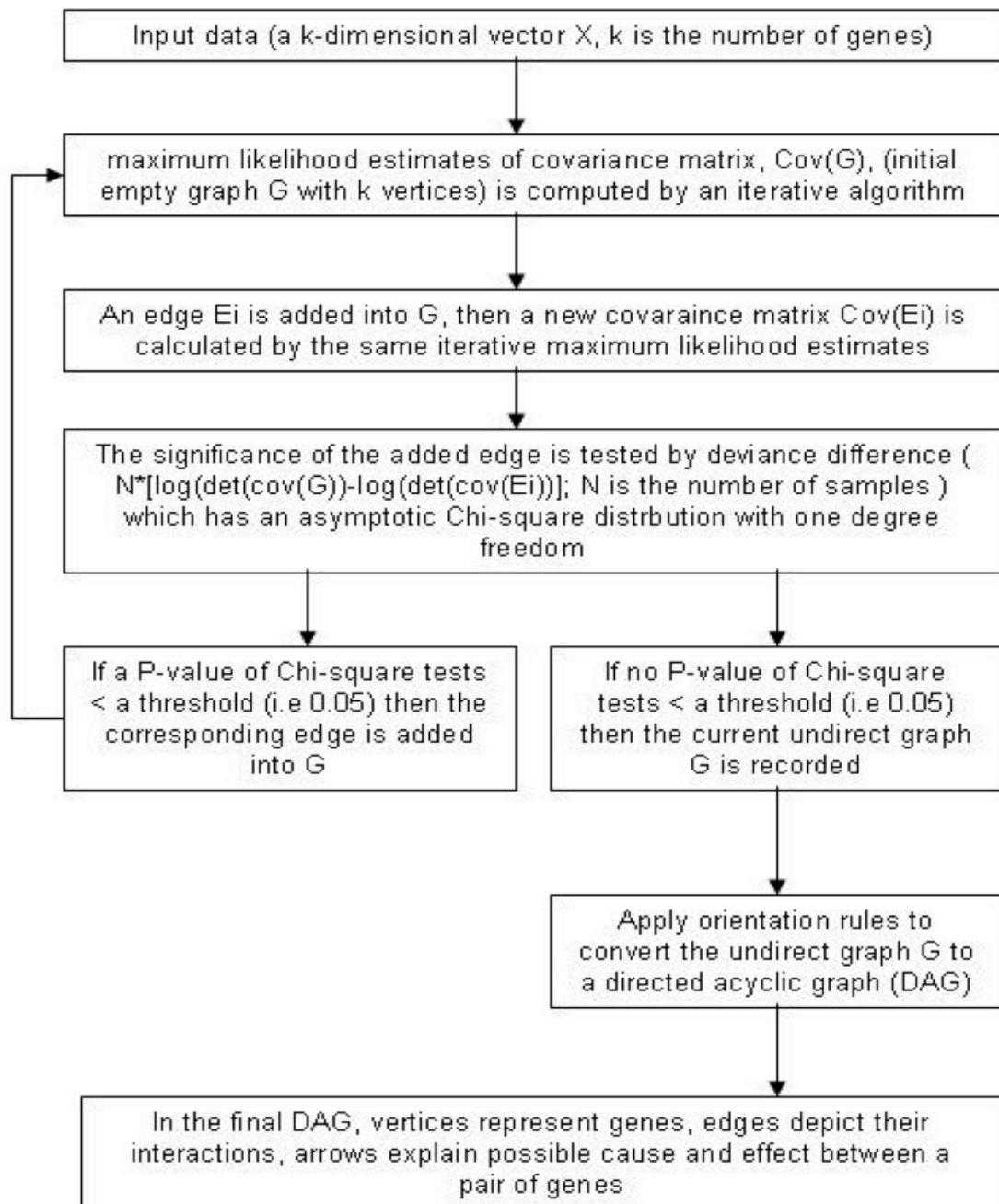


Figure 8.5: A workflow chart of Gaussian Graphical Model with Forward search algorithm (GGF).

Figure (8.8) is for the inferred network using a published method. Figure (8.9) shows the simulation of the network model in Figure (8.7). Finally Figure (8.10) provides the simulation of the merged network.

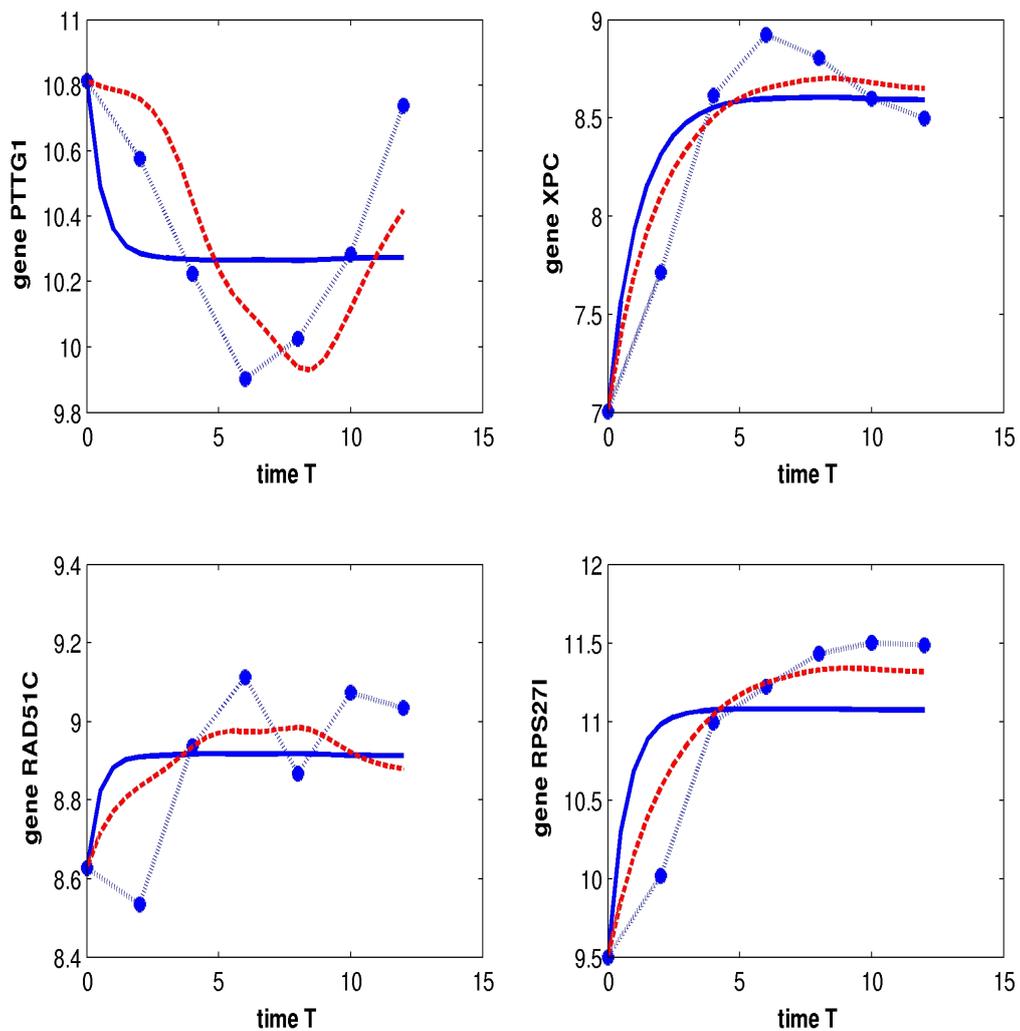


Figure 8.6: Simulations of the gene network model of eight genes: The dynamics of four genes were presented in the paper. Here are the remaining four genes: PTTG1, XPC, RAD51C, RPS27L (dash-star: microarray data; solid-line: simulation of the fully connected model; dash-line: simulation of the core network predicted by the GGF with eight mutual regulations).

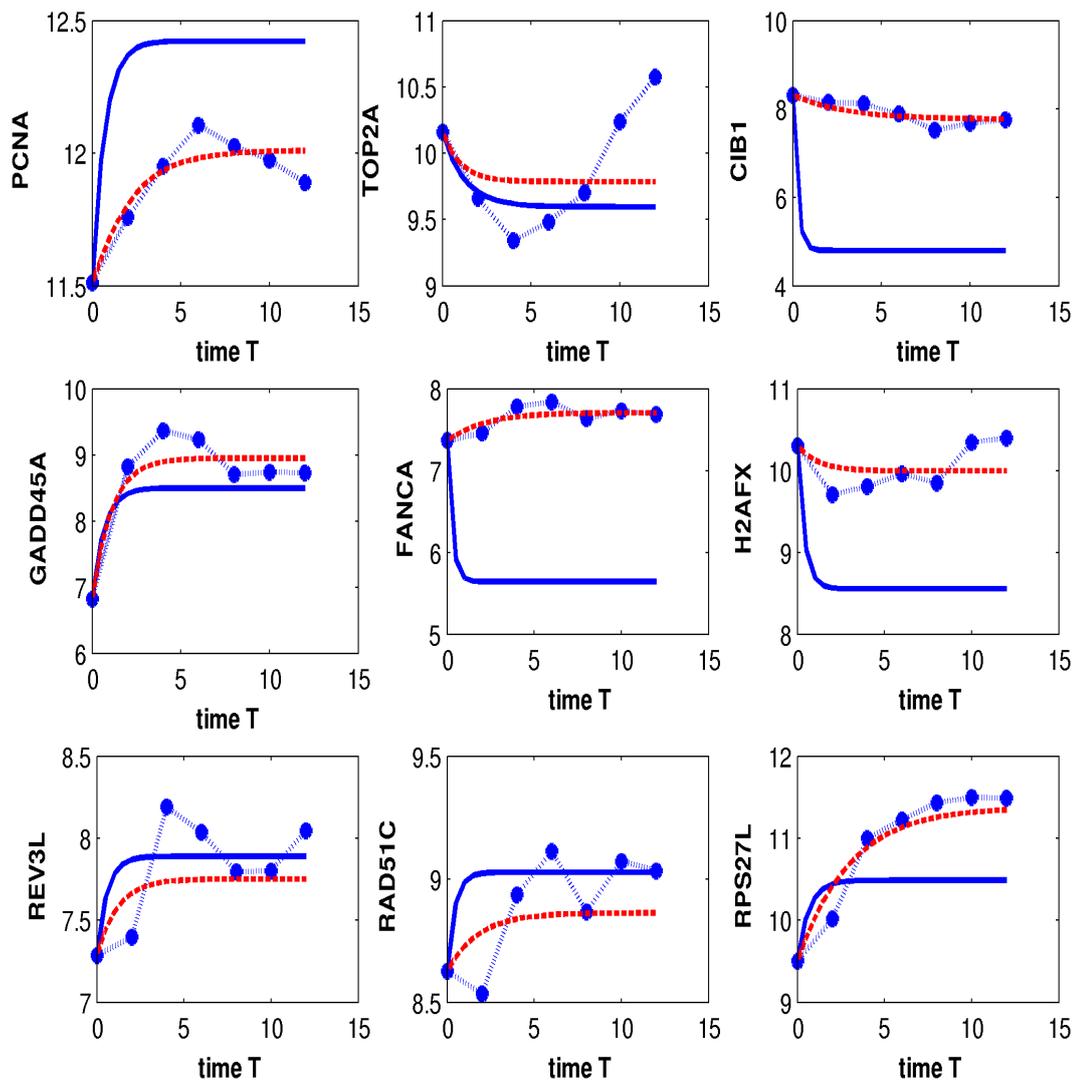


Figure 8.7: Simulations of the gene network of 21 genes: The dynamics of four genes of this network was presented in the paper in Figure (8.4). Here are the other 9 genes. (dash-star: microarray data; solid-line: simulation of the fully connected model; dash-line: simulation of the core network predicted by the GGF).

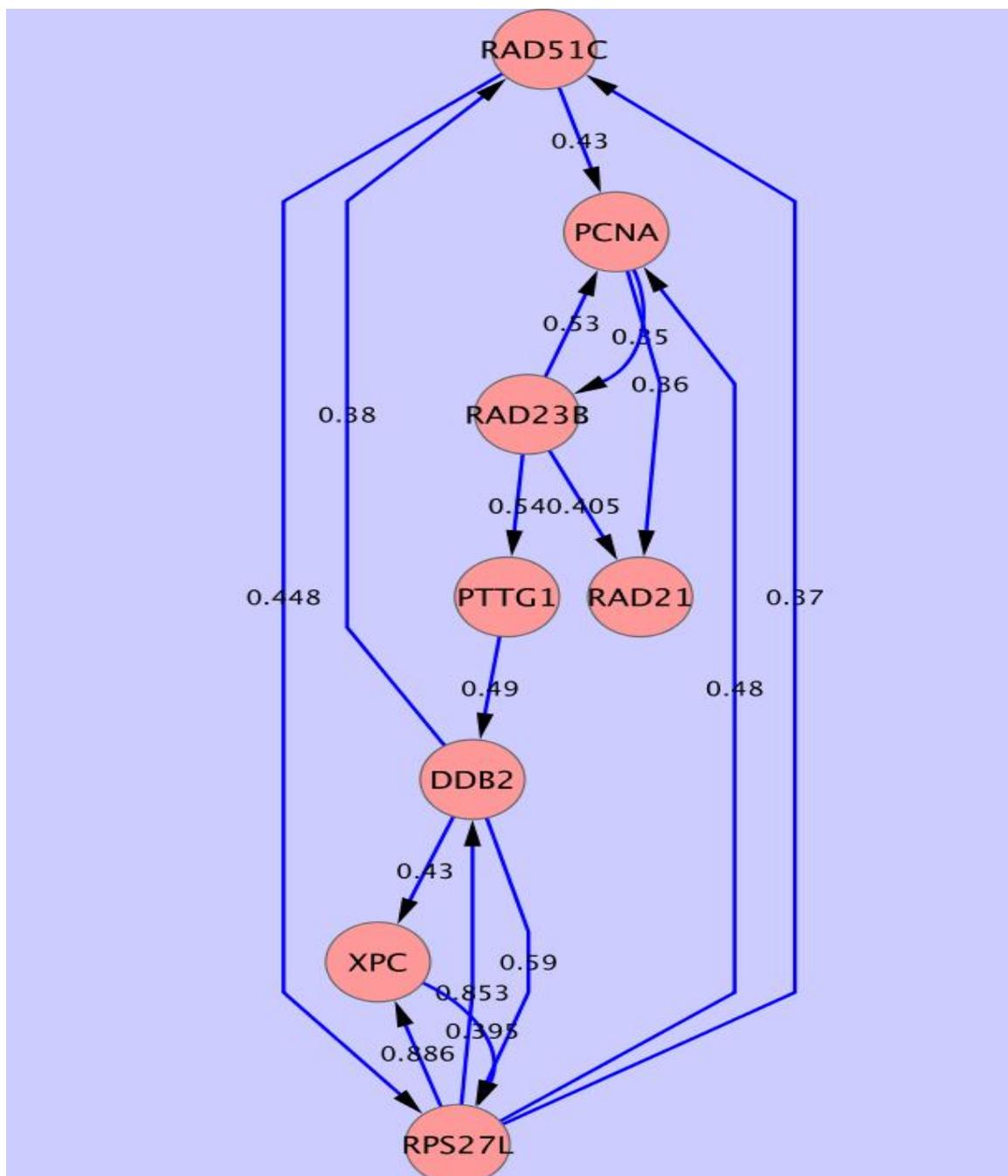


Figure 8.8: The graphic model of eight genes using a published inference method: Gene-gene interaction network was predicted by using the inference method in (Äijö and Lähdesmäki, 2009). The inferred network is a full matrix whose diagonal elements are zero. To match the inferred network in Fig. (8.2A), we selected the top 16 edges that have the largest values of the posterior probabilities.

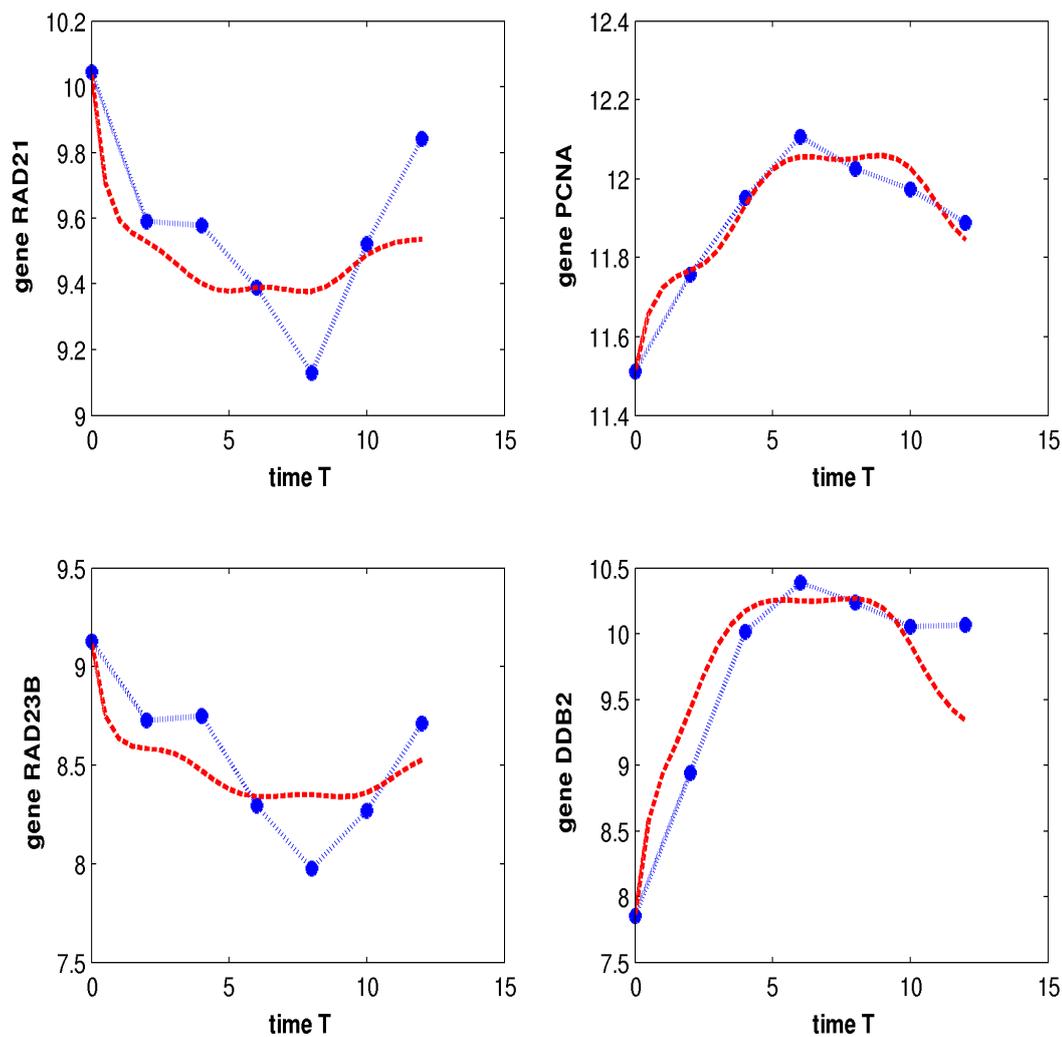


Figure 8.9: Simulation of the gene network model in Fig. (8.7): (Dash-star: microarray data; red-dash-line: simulation of the network model).

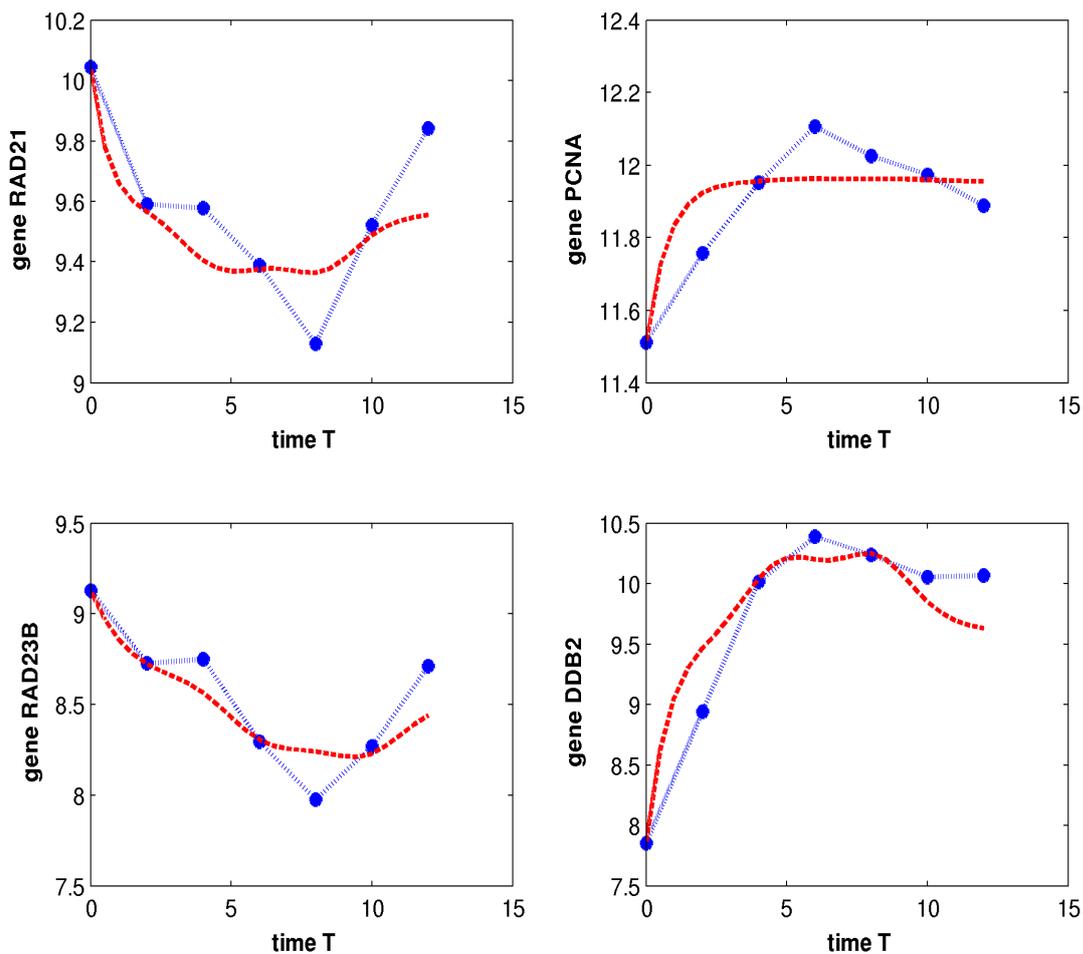


Figure 8.10: Simulation of the gene network of eight genes by merging the two networks in Fig. (8.2A) and Fig. (8.8) together: This merged network has 11 edges. (dash-star: microarray data; red-dash-line: simulation of the merged gene network).

Table 8.6: Model parameters of the fully connected network with 8 genes: Coefficient a_{ij}, b_{ij} and time delay τ_i in (8.2.3); Degradation rate d_i ; Basal transcription rate c_i ; Initial condition $x_i(0)$.

A								
16.3922	15.2941	9.2549	19.0588	12.5490	17.0196	19.5294	18.5882	0
8.5490	5.4118	9.5686	16.0000	8.7843	11.5294	18.9020	2.6667	17.8039
19.8431	19.2157	13.9608	6.7451	18.4314	13.9608	14.4314	19.2941	0
17.7255	19.5294	18.8235	19.8431	13.2549	17.1765	18.9804	19.9216	19.9216
0.7059	0.9686	0.6275	0.4549	0.3020	0.4627	0.8863	0.8627	0
16.3137	2.4314	17.8824	16.9412	9.1765	19.2941	18.3529	14.9804	18.5098
15.6078	18.1176	12.0784	0.6275	10.6667	8.3137	5.8039	17.9608	19.2941
2.2745	13.4902	13.9608	12.6275	3.5686	8.7059	16.3137	3.2941	17.3333
B								
5.0980	2.1176	2.7451	5.3333	9.6471	0.2353	0.7843	0.0784	19.8431
1.4118	12.1569	18.0392	6.8235	4.9412	13.3333	0.6275	12.0000	1.1765
3.4510	7.8431	13.0980	10.7451	5.8824	4.9412	6.1961	6.1176	18.7451
0.1569	0.6275	6.5098	0.1569	1.2549	0	0.0784	4.3922	0
1.2549	10.5098	1.4902	0.0784	1.4118	0.1569	3.2157	0.6275	19.5294
10.5098	16.0000	5.0196	14.1176	2.1176	8.7843	4.0000	17.0196	0.0784
16.3922	12.1569	14.0392	2.9804	7.4510	6.3529	5.4902	6.5882	6.1961
14.9804	8.6275	17.0196	8.9804	17.1765	7.5294	18.5098	12.6275	0.3137
τ_i								
	0	1.0700	1.3300	0	0	0	0	0
d_i								
	2.5098	1.1765	2.4314	2.5882	1.8039	0.8627	2.3529	1.4118
c_i								
	9.1373	12.9412	18.3529	14.4314	18.2745	6.0392	19.6863	14.9020
$x_i(0)$								
	10.0448	11.5111	9.1283	7.8532	10.8128	7.0051	8.6273	9.5000

Table 8.7: Model parameters of the core network with 8 genes in Fig. (8.1A): Coefficient a_{ij}, b_{ij} and time delay τ_i in (8.2.3); Degradation rate d_i ; Basal transcription rate c_i ; Initial condition $x_i(0)$.

A								
0	0	10.667	0	18.67	0	0	0	0
0	0	0	0	0	19.529	0	0	19.764
10.117	0	0	0	0	0	0	0	0
0	0	0	0	0	7.529	0	5.098	20.000
12.392	0	0	0	0	0	0	7.608	0
0	11.607	0	18.902	0	0	0	0	19.372
0	0	0	0	0	0	0	6.039	20.000
0	0	0	7.294	6.667	0	1.882	0	20.000
B								
0	0	12.392	0	1.176	0	0	0	20.000
0	0	0	0	0.862	7.843	0	0	10.352
4.078	0	0	0	0	0	0	0	10.117
0	0	0	0	0	9.803	0	7.764	0
3.686	0.235	0	0	0	0	0	1.882	10.509
0	10.039	0	12.235	0	0	0	0	0
0	0	0	0	0	0	0	10.274	0
0	0	0	6.902	16.3922	0	13.019	0	0.078
τ_i								
	0	1.0700	1.3300	0	0	0	0	0
d_i								
	0.6627	0.5059	0.5765	0.8157	0.3137	0.5647	0.9059	0.4627
c_i								
	4.0000	3.9216	2.5882	7.4510	0	3.5294	7.4510	4.7843
$x_i(0)$								
	10.0448	11.5111	9.1283	7.8532	10.8128	7.0051	8.6273	9.5000

8.6 Conclusion

In summary, this work proposed a new integrated approach that first uses a top-down method to infer the network structure, and then use a bottom-up approach to conduct detailed studies of gene regulatory networks. *In silico* experiments show that in addition to simulation error of dynamic model, AIC value, parameter identifiability and robustness property of regulatory networks are important criteria to select the optimal network from a variety of candidates. It is expected that our proposed new method will be used to design detailed mathematical models for investigating the dynamics of genetic regulation.

Chapter 9
Conclusion

Chapter 9

Conclusion

9.1 Contributions of the Thesis

In this thesis, the aim is to develop new methods for modelling and inference of biological systems such as gene regulatory networks. In the literature, researchers usually apply stochastic or deterministic methods to build the network structure and use Bayesian approach to estimate the unknown parameters. However, these approaches may not fully explain system dynamics efficiently. This thesis makes substantial contributions to computational biology by proposing new methodologies that can efficiently explain the system dynamics for particular biological systems. Through the application of these new methods, it is empirically shown that the new methods can yield realistic results and they appear to be able to provide insights into the understanding of biological networks .

In Chapter 2 and Chapter 3, a new model is proposed by which we can visualize multi-step chemical reactions system that is a fundamental issue in computational biology and bioinformatics. In order to simplify the system, a new concept (e.g. the length of a molecule) as an additional information is considered to

characterize system dynamics, which is defined as the location of a molecule in the multi-step reactions. An ODE model is used to find the optimal value in the non-linear function that represents the probability of the firing of last reaction in the system. To calibrate this probability function, a stochastic simulation method is proposed to calculate the probabilities under various system states. Numerical results suggest that this probability is dependent on the number of reaction steps but independent of the total molecule number, which leads to further development of a simplified model based on the network structure. Then our proposed two-variable model is applied to simulate the dynamics of mRNA degradation using experimentally observed data. Numerical results suggest that the length of molecules, which is approximately a half of the maximal length initially, played an important role in realizing experimental data.

In Chapter 4, the main contribution is to model multi-step chemical reaction events with time delays. Using both the analytical solution and stochastic simulation of the multi-step chemical reactions to obtain the relationship between the system state and value of time delay, a delay stochastic simulation algorithm is established. The proposed model is applied to model the degradation process of mRNA molecules based on experimental data in single cells for two separate systems. Our model both provides good accuracy for mRNA degradation as well as gene expression, which indicates that the proposed method is an effective approach to approximate multi-step reaction system more accurately. Half-life is an important concept to measure the degradation of species in biology. However, for many of the biological molecules, the decay process follows multi-step reactions rather than one-step reactions. Thus, the molecules at the intermediate states are also important for determining the value of species half-life. That may be the reason to explain the difference between the determined half-time under different experimental conditions. Using the inferred degradation rate in the

state-dependent delay model, our results suggest that our calculated half-time of mRNA molecules are between the determined values in the published papers.

In Chapter 5 and Chapter 6, we first study the ABC algorithm extensively tested on two chemical reaction systems. It's found that the ABC algorithm is an effective inference method that is capable of dealing with inference problems whose likelihood functions are hard to compute. Computational tests are conducted to examine the influence of a number of factors on the estimation error of the ABC algorithm. From that, it's noticed that taking different step sizes would not lead to distinct results. In addition, based on the framework of ABC SMC, two novel algorithms for the inference of unknown parameters in complex stochastic models for chemical reaction systems are proposed. These new algorithms impose stricter criteria to measure the simulation error, and the accuracy and efficiency are examined on two test problems. It is discovered that taking smaller values of discrepancy tolerance will result in more accurate estimates of unknown model parameters. This conclusion is confirmed by the second system that has been tested under different conditions. Numerical results suggest that the proposed new algorithms are promising methods to infer parameters in high-dimensional and complex biological system models and have better accuracy compared with the results of the published method. The encouraging result is that new algorithms do not need more computing time to achieve such accuracy.

In Chapter 7, we analyse sensitivity and robustness properties of biological systems simultaneously. In this framework, sensitivity analysis uses the difference method to calculate the derivatives of the probability density function; and robustness analysis is based on a general definition. Meanwhile a stochastic model of Nanog gene network with intrinsic noise based on a published model that discussed extrinsic noise only is proposed. Numerical results suggest that the system dynamics is sensitive to variations of one parameter that is related to the positive regulation from one complex. In addition, the change of a number of

other parameters will vary the bistability property of the Nanog gene network model. Numerical simulations also indicate that the proposed framework is an efficient approach to assess the robustness and sensitivity properties of biological network models.

In Chapter 8, the main contribution is the built of a novel approach for inferring genetic regulatory network combining both top-down and bottom-up approaches. To address issues regarding multiple regulations and sparseness of network topology, the probabilistic graphic models is used to infer network structure. By choosing various significant levels of the graphic models, a core network and an extended network are derived. Then a new mathematical model to represent complex regulation relationships is designed to validate the predicted graphic models and also to investigate detailed dynamic regulations. Computational results show that our predicted core network has smaller AIC values, parameter identifiability property and better robustness property. Subsequently, possibility of adding regulations to or removing regulations from the core network model are also tested, and AIC values and robustness property are selected as key criteria to validate the predicted models. Comparing with a published inference method, numerical results suggest that the proposed method could predict network models that have better simulation accuracy and robustness property. This study indicates that top-down approaches can provide initial simplified network structure, and more importantly, improve the accuracy and efficiency of the bottom-up approaches.

9.2 Future Directions

These results clearly demonstrate that the merits of which can outperform other investigated modelling and inference methods in the literature. However, the results also raise a number of questions which deserve further research.

1. *Mathematical modelling method.*

Regarding the two-variable modelling method, in order to make the method more convenient to use for studying other biological systems such as telomere length regulation, further research may involve the refinement of the probability function to increase the accuracy. On top of that, the current focus is limited to multi-step chemical reaction systems. Further work should investigate whether this two-variable modelling method is still valid for more complex systems.

Regarding the inference of genetic regulatory networks, in addition to using simulation error as the key criterion to select the optimal model, other measurements, such as AIC value, parameter identifiability and robustness property of a network, are also needed as important criteria. Meanwhile, a potential next step is to develop effective Bayesian algorithms for inferring regulatory networks using large-scale -omics datasets.

2. *Parameter inference method.*

Dealing with inference problems, a subjective choice of proper fitness tolerance values when performing the Bayesian inference makes the proposed method flexible, which can be chosen according to the accuracy we are interested in. Thus more sophisticated techniques, such as the adaptive selection methods, are needed to select the threshold values in the Approximate Bayesian Computation (ABC) algorithms. In addition, to reduce the computing time, more effective methods should be used to simulate the biological systems, such as the τ -leap methods and multi-scale simulation methods. Another alternative approach is to use parallel computing to reduce the heavy computing loads. All these issues are potential topics for future research work for parameter inference method.

In other aspects of inference method, sensitivity and robustness properties can be used to measure the variation of system dynamics caused by parameter perturbations. Sensitivity measures quantitative changes of variable values in the model; while robustness is the property of a system to maintain certain key properties, such as the bistability or oscillation. For the Nanog network model, results in this thesis raise a number of interesting questions regarding the sensitivity and robustness analysis, such as the quantitative definitions of these two properties and relationship between them. This deserves attention in the future.

Bibliography

Bibliography

- Adra S, Sun T, MacNeil S, Holcombe M, Smallwood R. 2010. Development of a three dimensional multiscale computational model of the human epidermis. *PloS one* **5**(1): e8511.
- Agrawal S, Archer C, Schaffer DV. 2009. Computational models of the notch network elucidate mechanisms of context-dependent signaling. *PLoS Comput Biol* **5**(5): e1000390, doi: 10.1371/journal.pcbi.1000390.
- Äijö T, Lähdesmäki H. 2009. Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics* **25**(22): 2937–2944.
- Akaike H. 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* **19**(6): 716–723.
- Akutsu T, Miyano S, Kuhara S. 2000. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics* **16**(8): 727–734.
- Ali M, Storey C, Törn A. 1997. Application of stochastic global optimization algorithms to practical problems. *Journal of Optimization Theory and Applications* **95**(3): 545–563.
- Almeida JS, Voit EO. 2003. Neural-network-based parameter estimation in s-system models of biological networks. *Genome Informatics* **14**: 114–123.

- Apri M, Molenaar J, De Gee M, Van Voorn G. 2010. Efficient estimation of the robustness region of biological models with oscillatory behavior. *PloS one* **5**(4): e9865.
- Ardehali MB, Lis JT. 2009. Tracking rates of transcription and splicing in vivo. *Nature structural & molecular biology* **16**(11): 1123–1124.
- Arkin A, Ross J, McAdams HH. 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ -infected escherichia coli cells. *Genetics* **149**(4): 1633–1648.
- Asher RB, Sebesta HR. 1971. Optimal control of systems with state-dependent time delay? *International Journal of Control* **14**(2): 353–366.
- Ashyraliyev M, Fomekong-Nanfack Y, Kaandorp JA, Blom JG. 2009. Systems biology: parameter estimation for biochemical models. *FEBS Journal* **276**(4): 886–902, doi: 10.1111/j.1742-4658.2008.06844.x, URL <http://dx.doi.org/10.1111/j.1742-4658.2008.06844.x>.
- Ashyraliyev M, Jaeger J, Blom J. 2008. Parameter estimation and determinability analysis applied to drosophila gap gene circuits. *BMC Systems Biology* **2**(1): 83, doi: 10.1186/1752-0509-2-83, URL <http://www.biomedcentral.com/1752-0509/2/83>.
- Bar-Joseph Z, Gerber G, Simon I, Gifford DK, Jaakkola TS. 2003. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proceedings of the National Academy of Sciences* **100**(18): 10146–10151.
- Bar-Joseph Z, Gitter A, Simon I. 2012. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics* **13**(8): 552–564.

- Bar-Yam Y. 1997. *Dynamics of complex systems*. Addison-Wesley studies in non-linearity, Addison-Wesley, ISBN 9780201557480, URL https://books.google.com.au/books?id=LH_wAAAAMAAJ.
- Barenco M, Tomescu D, Brewer D, Callard R, Stark J, Hubank M. 2006. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome biology* 7(3): R25.
- Barrio M, Burrage K, Leier A, Tian T. 2006. Oscillatory regulation of hes1: Discrete stochastic delay modelling and simulation. *PLoS Computational Biology* 2(9), URL <http://dblp.uni-trier.de/db/journals/ploscb/ploscb2.html#BarrioBLT06>.
- Bates D, Cosentino C. 2011. Validation and invalidation of systems biology models using robustness analysis. *Systems Biology, IET* 5(4): 229–244.
- Battogtokh D, Asch D, Case M, Arnold J, Schüttler HB. 2002. An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of neurospora crassa. *Proceedings of the National Academy of Sciences* 99(26): 16 904–16 909.
- Bayati B, Chatelain P, Koumoutsakos P. 2009. D-leaping: Accelerating stochastic simulation algorithms for reactions with delays. *Journal of Computational Physics* 228(16): 5908 – 5916, doi: 10.1016/j.jcp.2009.05.004, URL <http://www.sciencedirect.com/science/article/pii/S0021999109002435>.
- Beaumont MA, Cornuet JM, Marin JM, Robert CP. 2009. Adaptive approximate bayesian computation. *Biometrika* : asp052.
- Beaumont MA, Zhang W, Balding DJ. 2002. Approximate bayesian computation in population genetics. *Genetics* 162(4): 2025–2035.
- Blum MG, François O. 2010. Non-linear regression models for approximate bayesian computation. *Statistics and Computing* 20(1): 63–73.

- Boender CGE, Romeijn HE. 1995. Stochastic methods. In: *Handbook of global optimization*, Springer, pp. 829–869.
- Bokes P, King JR, Wood AT, Loose M. 2012. Multiscale stochastic modelling of gene expression. *Journal of mathematical biology* **65**(3): 493–520.
- Bornholdt S. 2005. Less is more in modeling large genetic networks. *SCIENCE-NEW YORK THEN WASHINGTON*- **310**(5747): 449.
- Bortz D, Nelson P. 2006. Model selection and mixed-effects modeling of hiv infection dynamics. *Bulletin of mathematical biology* **68**(8): 2005–2025.
- Boys RJ, Wilkinson DJ, Kirkwood TB. 2008. Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing* **18**(2): 125–135.
- Branz SE. 1996. A primer to mechanism in organic chemistry (sykes, peter). *Journal of Chemical Education* **73**(12): A313, doi: 10.1021/ed073pA313.2, URL <http://pubs.acs.org/doi/abs/10.1021/ed073pA313.2>.
- Bratsun D, Volfson D, Tsimring LS, Hasty J. 2005. Delay-induced stochastic oscillations in gene regulation. *Proceedings of the National Academy of Sciences of the United States of America* **102**(41): 14 593–14 598, doi: 10.1073/pnas.0503858102, URL <http://www.pnas.org/content/102/41/14593.abstract>.
- Bregman A, Avraham-Kelbert M, Barkai O, Duek L, Guterman A, Choder M. 2011. Promoter elements regulate cytoplasmic mrna decay. *Cell* **147**(7): 1473–83.
- Brett T, Galla T. 2013. Stochastic processes with distributed delays: chemical langevin equation and linear-noise approximation. *Physical review letters* **110**(25): 250 601.
- Brooks S. 1998. Markov chain monte carlo method and its application. *Journal of the royal statistical society: series D (the Statistician)* **47**(1): 69–100.

- Brooks SP, Morgan BJ. 1995. Optimization using simulated annealing. *The Statistician* : 241–257.
- Brown KS, Sethna JP. 2003. Statistical mechanical approaches to models with many poorly known parameters. *Physical Review E* **68**(2): 021 904.
- Bruggeman FJ, Westerhoff HV. 2007. The nature of systems biology. *TRENDS in Microbiology* **15**(1): 45–50.
- Burgess DJ. 2014. Synthetic biology: Cut up to bring together. *Nature Reviews Genetics* **15**(6): 365–365.
- Burrage K, Hancock J, Leier A, Jr DN. 2007. Modelling and simulation techniques for membrane biology. *Briefings in Bioinformatics* **8**(4): 234–244, URL <http://eprints.qut.edu.au/44378/>.
- Burrage K, Tian T, Burrage P. 2004. A multi-scaled approach for simulating chemical reaction systems. *Progress in biophysics and molecular biology* **85**(2): 217–234.
- Butler A, Glasbey C, Allcroft A, Wanless S. 2006. A latent gaussian model for compositional data with structural zeroes. (*Biomathematics and Statistics Scotland, Edinburgh*) technical report. .
- Cai L, Friedman N, Xie XS. 2006. Stochastic protein expression in individual cells at the single molecule level. *Nature* **440**(7082): 358–362.
- Cai X. 2007. Exact stochastic simulation of coupled chemical reactions with delays. *The Journal of chemical physics* **126**(12): 124 108.
- Cai X, Wang X. 2007. Stochastic modeling and simulation of gene networks—a review of the state-of-the-art research on stochastic simulations .
- Cao D, Parker R. 2001. Computational modeling of eukaryotic mRNA turnover. *RNA* **7**(9): 1192–1212, URL <http://rnajournal.cshlp.org/content/7/9/1192.abstract>.

- Cao D, Parker R. 2003. Computational modeling and experimental analysis of nonsense-mediated decay in yeast. *Cell* **113**(4): 533 – 545, doi: 10.1016/S0092-8674(03)00353-2, URL <http://www.sciencedirect.com/science/article/pii/S0092867403003532>.
- Cao J, Zhao H. 2008. Estimating dynamic models for gene regulation networks. *Bioinformatics* **24**(14): 1619–1624.
- Cao Y, Fan J, Gard TC. 1992. The effects of state-dependent time delay on a stage-structured population growth model. *Nonlinear Analysis: Theory, Methods & Applications* **19**(2): 95–105.
- Cao Y, Samuels DC. 2009. Discrete stochastic simulation methods for chemically reacting systems. *Methods in enzymology* **454**: 115–140.
- Chen SF, Juang YL, Chou WK, Lai JM, Huang CYF, Kao CY, Wang FS. 2009. Inferring a transcriptional regulatory network of the cytokinesis-related genes by network component analysis. *BMC systems biology* **3**(1): 110.
- Chickarmane V, Olariu V, Peterson C. 2012. Probing the role of stochasticity in a model of the embryonic stem cell–heterogeneous gene expression and reprogramming efficiency. *BMC systems biology* **6**(1): 98.
- Chickarmane V, Troein C, Nuber UA, Sauro HM, Peterson C. 2006. Transcriptional dynamics of the embryonic stem cell switch. *PLoS computational biology* **2**(9): e123.
- Chipperfield A, Fleming P, Fonseca C. 1994. Genetic algorithm tools for control systems engineering. In: *Proceedings of Adaptive Computing in Engineering Design and Control*. Citeseer, pp. 128–133.
- Chou IC, Voit EO. 2009. Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Mathematical biosciences* **219**(2): 57–83.

- Chubb JR, Liverpool TB. 2010. Bursts and pulses: insights from single cell studies into transcriptional mechanisms. *Current opinion in genetics & development* **20**(5): 478–484.
- Citri A, Yarden Y. 2006. Egf–erbb signalling: towards the systems level. *Nature reviews Molecular cell biology* **7**(7): 505–516.
- Cole DJ, Morgan BJ, Titterton D. 2010. Determining the parametric structure of models. *Mathematical biosciences* **228**(1): 16–30.
- Cox DR, Hinkley DV. 1979. *Theoretical statistics*. CRC Press.
- Cox J, Mann M. 2011. Quantitative, high-resolution proteomics for data-driven systems biology. *Annual review of biochemistry* **80**: 273–299.
- Csete ME, Doyle JC. 2002. Reverse engineering of biological complexity. *science* **295**(5560): 1664–1669.
- Csilléry K, Blum MG, Gaggiotti OE, François O. 2010. Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution* **25**(7): 410–418.
- Daigle B, Roh M, Petzold L, Niemi J. 2012. Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. *BMC Bioinformatics* **13**(1): 68, doi: 10.1186/1471-2105-13-68, URL <http://www.biomedcentral.com/1471-2105/13/68>.
- Damiani C, Filisetti A, Graudenzi A, Lecca P. 2013. Parameter sensitivity analysis of stochastic models: Application to catalytic reaction networks. *Computational biology and chemistry* **42**: 5–17.
- Davidich MI, Bornholdt S. 2008. Boolean network model predicts cell cycle sequence of fission yeast. *PloS one* **3**(2): e1672.
- De Jong H. 2002. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology* **9**(1): 67–103.

- De Smet R, Marchal K. 2010. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology* **8**(10): 717–729.
- Del Moral P, Doucet A, Jasra A. 2006. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(3): 411–436, doi: 10.1111/j.1467-9868.2006.00553.x, URL <http://dx.doi.org/10.1111/j.1467-9868.2006.00553.x>.
- Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* : 1–38.
- Deng Z, Tian T. 2014. A continuous approach for inferring parameters in mathematical models of regulatory networks. *BMC bioinformatics* **15**(1): 256.
- Dimitrova E, García-Puente LD, Hinkelmann F, Jarrah AS, Laubenbacher R, Stigler B, Stillman M, Vera-Licona P. 2011. Parameter estimation for boolean models of biological networks. *Theoretical Computer Science* **412**(26): 2816–2826.
- Erban R, Kevrekidis IG, Adalsteinsson D, Elston TC. 2006. Gene regulatory networks: A coarse-grained, equation-free approach to multiscale computation. *The Journal of chemical physics* **124**(8): 084106.
- Esposito WR, Floudas CA. 2000. Global optimization for the parameter estimation of differential-algebraic systems. *Industrial & Engineering Chemistry Research* **39**(5): 1291–1310.
- Fallahi-Sichani M, Flynn JL, Linderman JJ, Kirschner DE. 2012a. Differential risk of tuberculosis reactivation among anti-tnf therapies is due to drug binding kinetics and permeability. *The Journal of Immunology* **188**(7): 3169–3178.
- Fallahi-Sichani M, Kirschner DE, Linderman JJ. 2012b. Nf- κ b signaling dynamics play a key role in infection control in tuberculosis. *Frontiers in physiology* **3**.

- Fearnhead P, Prangle D. 2012. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(3): 419–474.
- Friedman N, Linial M, Nachman I, Pe'er D. 2000. Using bayesian networks to analyze expression data. *Journal of computational biology* **7**(3-4): 601–620.
- Gadkar K, Gunawan R, Doyle F. 2005. Iterative approach to model identification of biological networks. *BMC Bioinformatics* **6**(1): 155, doi: 10.1186/1471-2105-6-155, URL <http://www.biomedcentral.com/1471-2105/6/155>.
- Gandhi SJ, Zenklusen D, Lionnet T, Singer RH. 2011. Transcription of functionally related constitutive genes is not coordinated. *Nature structural & molecular biology* **18**(1): 27–34.
- Gardner TS, Di Bernardo D, Lorenz D, Collins JJ. 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**(5629): 102–105.
- Garneau NL, Wilusz J, Wilusz CJ. 2007. The highways and byways of mrna decay. *Nature reviews Molecular cell biology* **8**(2): 113–126.
- Gibson MA, Bruck J. 2000. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The journal of physical chemistry A* **104**(9): 1876–1889.
- Gillespie D. 2007. Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem* **58**.
- Gillespie DT. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics* **22**(4): 403–434.

- Gillespie DT. 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**(25): 2340–2361, URL http://pubs3.acs.org/acs/journals/doi/lookup?in_doi=10.1021/j100540a008.
- Gillespie DT. 1992. A rigorous derivation of the chemical master equation. *Physica A Statistical Mechanics and its Applications* **188**: 404–425, doi: 10.1016/0378-4371(92)90283-V.
- Gillespie DT. 2001. Approximate accelerated stochastic simulation of chemically reacting systems. *J Chem Phys* **115**: 1716–1733.
- Glass L, Kauffman SA. 1973. The logical analysis of continuous, non-linear biochemical control networks. *Journal of theoretical Biology* **39**(1): 103–129.
- Glauche I, Herberg M, Roeder I. 2010. Nanog variability and pluripotency regulation of embryonic stem cells-insights from a mathematical model analysis. *PLoS One* **5**(6): e11 238.
- Goel G, Chou IC, Voit EO. 2008. System estimation from metabolic time-series data. *Bioinformatics* **24**(21): 2505–2511, doi: 10.1093/bioinformatics/btn470, URL <http://bioinformatics.oxfordjournals.org/content/24/21/2505.abstract>.
- Golightly A, Wilkinson DJ. 2005. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics* **61**(3): 781–788.
- Golightly A, Wilkinson DJ. 2006. Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology* **13**(3): 838–851.
- Golightly A, Wilkinson DJ. 2011. Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface Focus* **1**(6): 807–820.

- Gonzalez OR, Küper C, Jung K, Naval PC, Mendoza E. 2007. Parameter estimation using simulated annealing for s-system models of biochemical networks. *Bioinformatics* **23**(4): 480–486, doi: 10.1093/bioinformatics/btl522, URL <http://dx.doi.org/10.1093/bioinformatics/btl522>.
- Greenstein JL, Winslow RL. 2011. Integrative systems models of cardiac excitation–contraction coupling. *Circulation research* **108**(1): 70–84.
- Grigull J, Mnaimneh S, Pootoolal J, Robinson MD, Hughes TR. 2004. Genome-wide analysis of mrna stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Molecular and cellular biology* **24**(12): 5534–5547.
- Grossmann IE. 1996. *Global optimization in engineering design*, vol. 9. Springer.
- Gunawan R, Cao Y, Petzold L, Doyle FJ. 2005. Sensitivity analysis of discrete stochastic systems. *Biophysical Journal* **88**(4): 2530–2540.
- Haefner JW. 2005. *Modeling biological systems:: Principles and applications*. Springer Science & Business Media.
- Hartig F, Calabrese JM, Reineking B, Wiegand T, Huth A. 2011. Statistical inference for stochastic simulation models–theory and application. *Ecology letters* **14**(8): 816–827.
- Hecker M, Lambeck S, Toepfer S, Van Someren E, Guthke R. 2009. Gene regulatory network inference: data integration in dynamic models-a review. *Biosystems* **96**(1): 86–103.
- Heitzler P, Simpson P. 1991. The choice of cell fate in the epidermis of drosophila. *Cell* **64**(6): 1083–1092.
- Herberg M, Kalkan T, Glauche I, Smith A, Roeder I. 2014. A model-based analysis of culture-dependent phenotypes of mescs. *PloS one* **9**(3): e92496.

- Hickman GJ, Hodgman TC. 2009. Inference of gene regulatory networks using boolean-network inference methods. *Journal of bioinformatics and computational biology* 7(06): 1013–1029.
- Hines KE, Middendorf TR, Aldrich RW. 2014. Determination of parameter identifiability in nonlinear biophysical models: A bayesian approach. *The Journal of general physiology* 143(3): 401–416.
- Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95(5): 717–728.
- Horst R, Tuy H. 1996. *Global optimization: Deterministic approaches*. Springer Science & Business Media.
- Hurn AS, Jeisman JI, Lindsay KA. 2007. Seeing the wood for the trees: A critical evaluation of methods to estimate the parameters of stochastic differential equations. *Journal of Financial Econometrics* 5(3): 390–455, doi: 10.1093/jjfinec/nbm009, URL <http://jfec.oxfordjournals.org/content/5/3/390.abstract>.
- Hurn AS, Lindsay KA. 1999. Estimating the parameters of stochastic differential equations. *Math. Comput. Simul.* 48(4-6): 373–384, doi: 10.1016/S0378-4754(99)00017-8, URL [http://dx.doi.org/10.1016/S0378-4754\(99\)00017-8](http://dx.doi.org/10.1016/S0378-4754(99)00017-8).
- Iba H, Mimura A. 2002. Inference of a gene regulatory network by means of interactive evolutionary computing. *Information Sciences* 145(3): 225–236.
- Johannes MS, Polson N. 2003. Mcmc methods for continuous-time financial econometrics. Available at SSRN 480461 .
- Jordan MI. 1998. *Learning in graphical models:[proceedings of the nato advanced study institute...: Ettore mairona center, erice, italy, september 27-october 7, 1996]*, vol. 89. Springer Science & Business Media.

- Kaern M, Elston T, Blake W, Collins J. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet* **6**(6): 451–64.
- Karlebach G, Shamir R. 2008. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology* **9**(10): 770–780.
- Kauffman SA. 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology* **22**(3): 437–467.
- Kaufmann BB, van Oudenaarden A. 2007. Stochastic gene expression: from single molecules to the proteome. *Current opinion in genetics & development* **17**(2): 107–112.
- Kikuchi S, Tominaga D, Arita M, Takahashi K, Tomita M. 2003. Dynamic modeling of genetic networks using genetic algorithm and s-system. *Bioinformatics* **19**(5): 643–650, doi: 10.1093/bioinformatics/btg027, URL <http://bioinformatics.oxfordjournals.org/content/19/5/643.abstract>.
- Kim SY, Imoto S, Miyano S. 2003. Inferring gene networks from time series microarray data using dynamic bayesian networks. *Briefings in bioinformatics* **4**(3): 228–235.
- Kimura S, Ide K, Kashihara A, Kano M, Hatakeyama M, Masui R, Nakagawa N, Yokoyama S, Kuramitsu S, Konagaya A. 2005. Inference of s-system models of genetic networks using a cooperative coevolutionary algorithm. *Bioinformatics* **21**(7): 1154–1163.
- Kiparissides A, Kucherenko S, Mantalaris A, Pistikopoulos E. 2009. Global sensitivity analysis challenges in biological systems modeling. *Industrial & Engineering Chemistry Research* **48**(15): 7168–7180.
- Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. 2012. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28**(24): 3290–3297.

- Kirkpatrick S, Gelatt CD, Vecchi MP. 1983. Optimization by simulated annealing. *Science* **220**(4598): 671–680, doi: 10.1126/science.220.4598.671.
- Kitano H. 2004. Biological robustness. *Nature Reviews Genetics* **5**(11): 826–837.
- Kitano H. 2007. Towards a theory of biological robustness. *Molecular systems biology* **3**(1): 137.
- Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H. 2008. *Systems biology in practice: concepts, implementation and application*. John Wiley & Sons.
- Kohn KW. 1999. Molecular interaction map of the mammalian cell cycle control and dna repair systems. *Molecular biology of the cell* **10**(8): 2703–2734.
- Kohn KW. 2001. Molecular interaction maps as information organizers and simulation guides. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **11**(1): 84–97.
- Kucherenko S, Iooss B. 2014. Derivative based global sensitivity measures. *arXiv preprint arXiv:1412.2619* .
- Kunath T, Saba-El-Leil MK, Almousailleakh M, Wray J, Meloche S, Smith A. 2007. Fgf stimulation of the erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. *Development* **134**(16): 2895–2902.
- Laise P, Di Patti F, Fanelli D, Masselli M, Arcangeli A. 2011. Deterministic and stochastic aspects of vegf-a production and the cooperative behavior of tumoral cell colony. *Journal of theoretical biology* **272**(1): 55–63.
- Lall R, Voit EO. 2005. Parameter estimation in modulated, unbranched reaction chains within biochemical systems. *Comput. Biol. Chem.* **29**(5): 309–318, doi: 10.1016/j.compbiolchem.2005.08.001, URL <http://dx.doi.org/10.1016/j.compbiolchem.2005.08.001>.

- Leier A, Marquez-Lago T, Burrage K. 2008. Generalized binomial tau-leap method for biochemical kinetics incorporating both delay and intrinsic noise. *J Chem Phys* **128**(20): 205 107.
- Leuenberger C, Wegmann D. 2010. Bayesian computation and model selection without likelihoods. *Genetics* **184**(1): 243–252.
- Lewis J. 2008. From signals to patterns: Space, time, and mathematics in developmental biology. *Science* **322**(5900): 399–403, doi: 10.1126/science.1166154, URL <http://www.sciencemag.org/content/322/5900/399.abstract>.
- Liang S, Fuhrman S, Somogyi R. 1998. Reveal, a general reverse engineering algorithm for inference of genetic network architectures .
- Lillacci G, Khammash M. 2010. Parameter estimation and model selection in computational biology. *PLoS Computational Biology* **6**(3), URL <http://dblp.uni-trier.de/db/journals/ploscb/ploscb6.html#LillacciK10>.
- Little MP, Heidenreich WF, Li G. 2010. Parameter identifiability and redundancy: theoretical considerations. *PloS one* **5**(1): e8915.
- Liu JS. 2008. *Monte carlo strategies in scientific computing*. Springer Science & Business Media.
- Liu PK, Wang FS. 2008. Inference of biochemical network models in s-system using multiobjective optimization approach. *Bioinformatics* **24**(8): 1085–1092.
- Lockhart DJ, Winzeler EA. 2000. Genomics, gene expression and dna arrays. *nature* **405**(6788): 827–836.
- Loy CJ, Lydall D, Surana U. 1999. Ndd1, a high-dosage suppressor ofcdc28-1n, is essential for expression of a subset of late-s-phase-specific genes in *saccharomyces cerevisiae*. *Molecular and cellular biology* **19**(5): 3312–3327.
- Ma J. 2010. *Gene expression and regulation*. Springer.

- Ma L, Wagner J, Rice JJ, Hu W, Levine AJ, Stolovitzky GA. 2005. A plausible model for the digital response of p53 to dna damage. *Proc Natl Acad Sci U S A* **102**(40): 14 266–71, URL <http://www.biomedsearch.com/nih/plausible-model-digital-response-p53/16186499.html>.
- Ma S, Gong Q, Bohnert HJ. 2007. An arabidopsis gene network based on the graphical gaussian model. *Genome research* **17**(11): 1614–1625.
- Mackey MC, Santillán M, Tyran-Kamińska M, Zeron ES. 2014. The utility of simple mathematical models in understanding gene regulatory dynamics. *In silico biology* .
- Maetschke SR, Madhamshettiwar PB, Davis MJ, Ragan MA. 2013. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in bioinformatics* : bbt034.
- Marino S, Hogue IB, Ray CJ, Kirschner DE. 2008. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of theoretical biology* **254**(1): 178–196.
- Marjoram P, Molitor J, Plagnol V, Tavaré S. 2003. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences* **100**(26): 15 324–15 328.
- Marjoram P, Tavaré S. 2006. Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics* **7**(10): 759–770.
- Marquez-Lago T, Leier A, Burrage K. 2010. Probability distributed time delays: integrating spatial effects into temporal models. *BMC Systems Biology* **4**: 1–16, URL <http://eprints.qut.edu.au/42737/>.
- Marquez-Lago T, Stelling J. 2010. Counter-intuitive stochastic behavior of simple gene circuits with negative feedback. *Biophys J* **98**(9): 1742–50.

- Masel J, Siegal ML. 2009. Robustness: mechanisms and consequences. *Trends in Genetics* **25**(9): 395–403.
- McAdams HH, Arkin A. 1997. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences* **94**(3): 814–819.
- McAdams HH, Arkin A. 1999. It's a noisy business! genetic regulation at the nanomolar scale. *Trends in genetics* **15**(2): 65–69.
- McQuarrie DA. 1967. Stochastic Approach to Chemical Kinetics. *Journal of Applied Probability* **4**(3): 413–478, doi: 10.2307/3212214, URL <http://dx.doi.org/10.2307/3212214>.
- Mendes P, Kell D. 1998. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* **14**(10): 869–883.
- Mier-y Terán-Romero L, Silber M, Hatzimanikatis V. 2010. The origins of time-delay in template biopolymerization processes. *PLoS Computational Biology* **6**(4), URL <http://dblp.uni-trier.de/db/journals/ploscb/ploscb6.html#Mier-y-Teran-RomeroSH10>.
- Mitchell P, Tollervey D. 2001. mrna turnover. *Current opinion in cell biology* **13**(3): 320–325.
- Moles CG, Mendes P, Banga JR. 2003. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome research* **13**(11): 2467–2474.
- Monk NA. 2003. Oscillatory expression of hes1, p53, and nf- κ b driven by transcriptional time delays. *Current Biology* **13**(16): 1409–1413.
- Müller T, Faller D, Timmer J, Swameye I, Sandra O, Klingmüller U. 2004. Tests for cycling in a signalling pathway. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **53**(4): 557–568.

- Murphy K, Mian S, *et al.* 1999. Modelling gene expression data using dynamic bayesian networks. Technical report, Technical report, Computer Science Division, University of California, Berkeley, CA.
- Navarro P, Festuccia N, Colby D, Gagliardi A, Mullin NP, Zhang W, Karwacki-Neisius V, Osorno R, Kelly D, Robertson M, *et al.* 2012. Oct4/sox2-independent nanog autorepression modulates heterogeneous nanog gene expression in mouse es cells. *The EMBO journal* **31**(24): 4547–4562.
- Nayak RR, Kearns M, Spielman RS, Cheung VG. 2009. Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome research* **19**(11): 1953–1962.
- Omony J. 2014. Biological network inference: a review of methods and assessment of tools and techniques. *Ann Res Rev Biol* **4**: 577–601.
- Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A. 2002. Regulation of noise in the expression of a single gene. *Nature genetics* **31**(1): 69–73.
- Padovan-Merhar O, Raj A. 2013. Using variability in gene expression as a tool for studying gene regulation. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **5**(6): 751–759.
- Pahle J. 2009. Biochemical simulations: stochastic, approximate stochastic and hybrid approaches. *Brief Bioinform* .
- Pal R, Bhattacharya S, Caglar MU. 2012. Robust approaches for genetic regulatory network modeling and intervention: a review of recent advances. *Signal Processing Magazine, IEEE* **29**(1): 66–76.
- Papamichail I, Adjiman CS. 2002. A rigorous global optimization algorithm for problems with ordinary differential equations. *Journal of Global Optimization* **24**(1): 1–33.

- Passos DO, Parker R. 2008. Analysis of cytoplasmic mrna decay in *saccharomyces cerevisiae*. *Methods in enzymology* **448**: 409–427.
- Paulsson J. 2004. Summing up the noise in gene networks. *Nature* **427**(6973): 415–418.
- Paulsson J. 2005. Models of stochastic gene expression. *Physics of life reviews* **2**(2): 157–175.
- Peleg M, Rubin D, Altman RB. 2005. Using petri net tools to study properties and dynamics of biological systems. *Journal of the American Medical Informatics Association* **12**(2): 181–199.
- Penfold CA, Wild DL. 2011. How to infer gene networks from expression profiles, revisited. *Interface Focus* **1**(6): 857–870.
- Picchini UL. 2014. Inference for sde models via approximate bayesian computation. *Journal of Computational and Graphical Statistics in press* .
- Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. 1999. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular Biology and Evolution* **16**(12): 1791–1798.
- Qin F, Li L. 2004. Model-based fitting of single-channel dwell-time distributions. *Biophysical Journal* **87**(3): 1657 – 1671, doi: 10.1021/ed073pA313.2, URL <http://www.sciencedirect.com/science/article/pii/S0006349504736474>.
- Quo CF, Moffitt RA, Merrill Jr AH, Wang MD. 2011. Adaptive control model reveals systematic feedback and key molecules in metabolic pathway regulation. *Journal of Computational Biology* **18**(2): 169–182.
- Raj A, van Oudenaarden A. 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**(2): 216–226.

- Raj A, van Oudenaarden A. 2009. Single-molecule approaches to stochastic gene expression. *Annual review of biophysics* **38**: 255.
- Rao S, van der Schaft A, van Eunen K, Bakker BM, Jayawardhana B. 2014. A model reduction method for biochemical reaction networks. *BMC systems biology* **8**(1): 52.
- Raser JM, O'Shea EK. 2004. Control of stochasticity in eukaryotic gene expression. *Science* **304**(5678): 1811–1814.
- Rathinam M, Sheppard PW, Khammash M. 2010. Efficient computation of parameter sensitivities of discrete stochastic chemical reaction networks. *The Journal of chemical physics* **132**(3): 034 103.
- Raue A, Karlsson J, Saccomani MP, Jirstrand M, Timmer J. 2014. Comparison of approaches for parameter identifiability analysis of biological systems. *Bioinformatics* : btt006.
- Reinker S, Altman R, Timmer J. 2006. Parameter estimation in stochastic biochemical reactions. *IEE Proceedings-Systems Biology* **153**(4): 168–178.
- Ribeiro AS. 2010. Stochastic and delayed stochastic models of gene expression and regulation. *Mathematical Biosciences* **223**(1): 1–11.
- Robert C, Casella G. 2013. *Monte carlo statistical methods*. Springer Science & Business Media.
- Rodda DJ, Chew JL, Lim LH, Loh YH, Wang B, Ng HH, Robson P. 2005. Transcriptional regulation of nanog by oct4 and sox2. *Journal of Biological Chemistry* **280**(26): 24 731–24 737.
- Rodriguez-Fernandez M, Mendes P, Banga JR. 2006. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems* **83**(2): 248–265.

- Roussel M, Zhu R. 2006. Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. *Phys Biol* **3**(4): 274–84.
- Rung J, Brazma A. 2013. Reuse of public genome-wide gene expression data. *Nature Reviews Genetics* **14**(2): 89–99.
- Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S. 2008. *Global sensitivity analysis: the primer*. John Wiley & Sons.
- Savageau MA. 1969. Biochemical systems analysis: I. some mathematical properties of the rate law for the component enzymatic reactions. *Journal of theoretical biology* **25**(3): 365–369.
- Scheff JD, Mavroudis PD, Calvano SE, Lowry SF, Androulakis IP. 2011. Modeling autonomic regulation of cardiac function and heart rate variability in human endotoxemia. *Physiological genomics* **43**(16): 951–964.
- Schlicht R, Winkler G. 2008. A delay stochastic process with applications in molecular biology. *Journal of Mathematical Biology* **57**: 613–648, doi: 10.1007/s00285-008-0178-y, URL <http://dx.doi.org/10.1007/s00285-008-0178-y>.
- Schlitt T, Brazma A. 2007. Current approaches to gene regulatory network modelling. *BMC bioinformatics* **8**(Suppl 6): S9.
- Schnell S. 2014. Validity of the michaelis–menten equation–steady-state or reactant stationary assumption: that is the question. *FEBS Journal* **281**(2): 464–472.
- Shi W, Wang H, Pan G, Geng Y, Guo Y, Pei D. 2006. Regulation of the pluripotency marker rex-1 by nanog and sox2. *Journal of Biological Chemistry* **281**(33): 23 319–23 325.
- Shyu AB, Wilkinson MF, Van Hoof A. 2008. Messenger RNA regulation: to translate or to degrade. *EMBO J* **2**(3): 471–478, doi: 10.1038/sj.emboj.7601977, URL <http://dx.doi.org/10.1038/sj.emboj.7601977>.

- Silva J, Nichols J, Theunissen TW, Guo G, van Oosten AL, Barrandon O, Wray J, Yamanaka S, Chambers I, Smith A. 2009. Nanog is the gateway to the pluripotent ground state. *Cell* **138**(4): 722–737.
- Simon R. 2008. Microarray-based expression profiling and informatics. *Current opinion in biotechnology* **19**(1): 26–29.
- Singer AB, Bok JK, Bartona PI. 2001. Convex underestimators for variational and optimal control problems. *Computer Aided Chemical Engineering* **9**: 767–772.
- Singhania R, Sramkoski RM, Jacobberger JW, Tyson JJ. 2011. A hybrid model of mammalian cell cycle regulation. *PLoS computational biology* **7**(2): e1001077.
- Sisson SA, Fan Y, Tanaka MM. 2007. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences* **104**(6): 1760–1765, doi: 10.1073/pnas.0607208104, URL <http://www.pnas.org/content/104/6/1760.abstract>.
- Small M. 2012. *Dynamics of biological systems / michael small*. CRC Press Boca Raton, Fla, ISBN 9781439853368.
- Spiller DG, Wood CD, Rand DA, White MR. 2010. Measurement of single-cell dynamics. *Nature* **465**(7299): 736–745.
- Spirtes P, Glymour C, Scheines R, Kauffman S, Aimale V, Wimberly F. 2000. Constructing bayesian network models of gene expression networks from microarray data .
- Srinivas M, Patnaik LM. 1994. Genetic algorithms: A survey. *Computer* **27**(6): 17–26.
- Srivastava R, Haseltine EL, Mastny E, Rawlings JB. 2011. The stochastic quasi-steady-state assumption: reducing the model but not the noise. *J Chem Phys* **134**(15): 154109, URL <http://www.biomedsearch.com/nih/stochastic-quasi-steady-state-assumption/21513377.html>.

- Steggles LJ, Banks R, Shaw O, Wipat A. 2007. Qualitatively modelling and analysing genetic regulatory networks: a petri net approach. *Bioinformatics* **23**(3): 336–343.
- Symonds MR, Moussalli A. 2011. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using akaike's information criterion. *Behavioral Ecology and Sociobiology* **65**(1): 13–21.
- Szallasi Z, Stelling J, Periwal V. 2006. Systems modeling in cell biology, from concepts to nuts and bolts.
- Tanaka MM, Francis AR, Luciani F, Sisson SA. 2006. Using approximate bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics* **173**(3): 1511–1520, doi: 10.1534/genetics.106.055574, URL <http://www.genetics.org/content/173/3/1511.abstract>.
- Tavaré S, Balding DJ, Griffiths RC, Donnelly P. 1997. Inferring coalescence times from dna sequence data. *Genetics* **145**(2): 505–518.
- Tegner J, Yeung MS, Hasty J, Collins JJ. 2003. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proceedings of the National Academy of Sciences* **100**(10): 5944–5949.
- Thomas P, Straube A, Grima R. 2012. The slow-scale linear noise approximation: an accurate, reduced stochastic description of biochemical networks under timescale separation conditions. *BMC Syst Biol* **6**(1): 39.
- Thomas R. 1973. Boolean formalization of genetic control circuits. *Journal of theoretical biology* **42**(3): 563–585.
- Thomas R, Paredes CJ, Mehrotra S, Hatzimanikatis V, Papoutsakis ET. 2007. A model-based optimization framework for the inference of regulatory interactions using time-course dna microarray expression data. *BMC bioinformatics* **8**(1): 228.

- Thornton K, Andolfatto P. 2006. Approximate bayesian inference reveals evidence for a recent, severe bottleneck in a netherlands population of drosophila melanogaster. *Genetics* **172**(3): 1607–1619, doi: 10.1534/genetics.105.048223, URL <http://www.genetics.org/content/172/3/1607.abstract>.
- Tian T. 2010. Stochastic models for inferring genetic regulation from microarray gene expression data. *Biosystems* **99**(3): 192–200.
- Tian T. 2013. Chemical memory reactions induced bursting dynamics in gene expression. *PLoS One* **8**(1): e52 029.
- Tian T. 2014. Simplified stochastic models with time delay for studying the degradation process of mRNA molecules. *Int. J. Data Min. Bioinformatics* **10**(1): 18–32, doi: 10.1504/IJDMB.2014.062891, URL <http://dx.doi.org/10.1504/IJDMB.2014.062891>.
- Tian T, Burrage K. 2004. Binomial leap methods for simulating stochastic chemical kinetics. *J Chem Phys* **121**(21): 10 356–64.
- Tian T, Burrage K, Burrage PM, Carletti M. 2007a. Stochastic delay differential equations for genetic regulatory networks. *Journal of Computational and Applied Mathematics* **205**(2): 696 – 707, doi: 10.1016/j.cam.2006.02.063, URL <http://www.sciencedirect.com/science/article/pii/S0377042706003943>.
- Tian T, Olson S, Whitacre JM, Harding A. 2011. The origins of cancer robustness and evolvability. *Integrative Biology* **3**(1): 17–30.
- Tian T, Smith-Miles K. 2014. Mathematical modelling of gata-switching for regulating the differentiation of hematopoietic stem cell. *BMC bioinformatics* **8**(S8): S8.
- Tian T, Song J. 2012. Mathematical modelling of the map kinase pathway using proteomic datasets. *PloS one* **7**(8): e42 230.

- Tian T, Xu S, Gao J, Burrage K. 2007b. Simulated maximum likelihood method for estimating kinetic rates in gene expression. *Bioinformatics* **23**(1): 84–91, doi: 10.1093/bioinformatics/btl552, URL <http://bioinformatics.oxfordjournals.org/content/23/1/84.abstract>.
- Tomlin CJ, Axelrod JD. 2007. Biology by numbers: mathematical modelling in developmental biology. *Nat. Rev. Genet* **8**(5): 331–340, doi: 10.1038/nrg2098, URL <http://dx.doi.org/10.1038/nrg2098>.
- Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf M. 2009. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface* **6**(31): 187–202.
- Törn A, Ali MM, Viitanen S. 1999. Stochastic global optimization: Problem classes and solution techniques. *Journal of Global Optimization* **14**(4): 437–447.
- Toyooka Y, Shimosato D, Murakami K, Takahashi K, Niwa H. 2008. Identification and characterization of subpopulations in undifferentiated es cell culture. *Development* **135**(5): 909–918.
- Trcek T, Larson DR, Moldón A, Query CC, Singer RH. 2011. Single-molecule mrna decay measurements reveal promoter-regulated mrna stability in yeast. *Cell* **147**(7): 1484–1497.
- Tsai KY, Wang FS. 2005. Evolutionary optimization with data collocation for reverse engineering of biological networks. *Bioinformatics* **21**(7): 1180–1188, doi: 10.1093/bioinformatics/bti099, URL <http://bioinformatics.oxfordjournals.org/content/21/7/1180.abstract>.
- Turner BM, Zandt TV. 2012. A tutorial on approximate bayesian computation. *Journal of Mathematical Psychology* **56**(2): 69 – 85, doi: 10.1016/j.jmp.2012.02.005, URL <http://www.sciencedirect.com/science/article/pii/S0022249612000272>.

- van Hoof A, Parker R. 2002. Messenger rna degradation: Beginning at the end. *Current Biology* **12**(8): R285 – R287, doi: 10.1016/S0960-9822(02)00802-3, URL <http://www.sciencedirect.com/science/article/pii/S0960982202008023>.
- Veis J, Klug H, Koranda M, Ammerer G. 2007. Activation of the g2/m-specific gene *clb2* requires multiple cell cycle signals. *Molecular and cellular biology* **27**(23): 8364–8373.
- Vilela M, Chou IC, Vinga S, Vasconcelos AT, Voit EO, Almeida JS. 2008. Parameter optimization in s-system models. *BMC systems biology* **2**(1): 35.
- Walpole J, Papin JA, Peirce SM. 2013. Multiscale computational models of complex biological systems. *Annual review of biomedical engineering* **15**: 137.
- Wang J. 2008. Computational biology of genome expression and regulation? a review of microarray bioinformatics. *Journal of Environmental Pathology, Toxicology and Oncology* **27**(3).
- Wang J, Bø TH, Jonassen I, Myklebost O, Hovig E. 2003a. Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. *BMC bioinformatics* **4**(1): 60.
- Wang J, Cheung LWK, Delabie J. 2005. New probabilistic graphical models for genetic regulatory networks studies. *Journal of biomedical informatics* **38**(6): 443–455.
- Wang J, Levasseur DN, Orkin SH. 2008. Requirement of nanog dimerization for stem cell self-renewal and pluripotency. *Proceedings of the National Academy of Sciences* **105**(17): 6326–6331.
- Wang J, Myklebost O, Hovig E. 2003b. Mgraph: graphical models for microarray data analysis. *Bioinformatics* **19**(17): 2210–2211.
- Wang J, Tian T. 2010. Quantitative model for inferring dynamic regulation of the tumour suppressor gene p53. *BMC bioinformatics* **11**(1): 36.

- Wang J, Wu Q, Tian T. 2015. An integrated approach to infer dynamic protein-gene interactions: a case study of the human p53 protein (submitted for publication)
- Wang SQ, Li HX. 2012. Bayesian inference based modelling for gene transcriptional dynamics by integrating multiple source of knowledge. *BMC systems biology* **6**(Suppl 1): S3.
- Wang Y, Chen J, Li Q, Wang H, Liu G, Jing Q, Shen B. 2011. Identifying novel prostate cancer associated pathways based on integrative microarray data analysis. *Computational Biology and Chemistry* **35**(3): 151–158.
- Wang Y, Christley S, Mjolsness E, Xie X. 2010. Parameter inference for discretely observed stochastic kinetic models using stochastic gradient descent. *BMC Systems Biology* **4**(1): 99, doi: 10.1186/1752-0509-4-99, URL <http://www.biomedcentral.com/1752-0509/4/99>.
- Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO. 2002. Precision and functional specificity in mrna decay. *Proceedings of the National Academy of Sciences* **99**(9): 5860–5865.
- Wegmann D, Leuenberger C, Excoffier L. 2009. Efficient approximate bayesian computation coupled with markov chain monte carlo without likelihood. *Genetics* **182**(4): 1207–1218.
- Wilkinson D. 2009. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat Rev Genet* **10**(2): 122–33.
- Wilkinson DJ. 2007. Bayesian methods in bioinformatics and computational systems biology. *Briefings in bioinformatics* **8**(2): 109–116.
- Wilkinson DJ. 2011. *Stochastic modelling for systems biology*. CRC press.

- Wittmann DM, Blöchl F, Trümbach D, Wurst W, Prakash N, Theis FJ. 2009. Spatial analysis of expression patterns predicts genetic interactions at the mid-hindbrain boundary. *PLoS computational biology* 5(11): e1000569.
- Wu Q, Jiang F, Tian T. 2015. Sensitivity and robustness analysis for stochastic model of nanog gene regulatory network. *International Journal of Bifurcation and Chaos* 25(07): 1540009, doi: 10.1142/S021812741540009X, URL <http://www.worldscientific.com/doi/abs/10.1142/S021812741540009X>.
- Wu Q, Smith-Miles K, Tian T. 2012. A two-variable model for stochastic modelling of chemical events with multi-step reactions. In: *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*. pp. 1–6, doi: 10.1109/BIBM.2012.6392681.
- Wu Q, Smith-Miles K, Tian T. 2013a. Approximate bayesian computation for estimating rate constants in biochemical reaction systems. In: *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*. pp. 416–421, doi: 10.1109/BIBM.2013.6732528.
- Wu Q, Smith-Miles K, Tian T. 2014. Approximate bayesian computation schemes for parameter inference of discrete stochastic models using simulated likelihood density. *BMC bioinformatics* 15(S12): S3, doi: 10.1186/1471-2105-15-S12-S3, URL <http://www.biomedcentral.com/1471-2105/15/S12/S3>.
- Wu Q, Smith-Miles K, Zhou T, Tian T. 2013b. Stochastic modelling of biochemical systems of multi-step reactions using a simplified two-variable model. *BMC Systems Biology* 7(4): S14, doi: 10.1186/1752-0509-7-S4-S14, URL <http://dx.doi.org/10.1186/1752-0509-7-S4-S14>.
- Wu Q, Tian T. 2015. Stochastic modelling of regulatory networks using state-dependent time delay (to be submitted) .

- Yeung MS, Tegnér J, Collins JJ. 2002. Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences* **99**(9): 6163–6168.
- Zhan C, Yeung LF. 2011. Parameter estimation in systems biology models using spline approximation. *BMC systems biology* **5**(1): 14.
- Zhang X, Liu K, Liu ZP, Duval B, Richer JM, Zhao XM, Hao JK, Chen L. 2013. Narromi: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics* **29**(1): 106–113.
- Zhou Y, Zhuang X. 2007. Kinetic analysis of sequential multistep reactions. *The Journal of Physical Chemistry B* **111**(48): 13 600–13 610, doi: 10.1021/jp073708+, URL <http://pubs.acs.org/doi/abs/10.1021/jp073708%2B>.
- Zhu R, Ribeiro AS, Salahub D, Kauffman SA. 2007. Studying genetic regulatory networks at the molecular level: Delayed reaction stochastic models. *Journal of Theoretical Biology* **246**(4): 725 – 745, doi: 10.1016/j.jtbi.2007.01.021, URL <http://www.sciencedirect.com/science/article/pii/S0022519307000513>.
- Zhu Y, Pan W, Shen X. 2009. Support vector machines with disease-gene-centric network penalty for high dimensional microarray data. *Statistics and its interface* **2**(3): 257.
- Zou M, Conzen SD. 2005. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* **21**(1): 71–79.