

MODELING AND IDENTIFICATION OF BIOMOLECULAR SYSTEMS

ABHISHEK DEY



**DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
MAY 2019**

© Indian Institute of Technology Delhi (IITD), New Delhi, 2019

MODELING AND IDENTIFICATION OF BIOMOLECULAR SYSTEMS

by

ABHISHEK DEY

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

in fulfilment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

MAY 2019

Certificate

This is to certify that the thesis entitled “**Modeling and Identification of Biomolecular Systems**”, submitted by **Abhishek Dey** to the Indian Institute of Technology Delhi, for the award of the degree of **Doctor of Philosophy** in Electrical Engineering, is a record of the original, bonafide research work carried out by him under my supervision and guidance. The thesis has reached the standards fulfilling the requirements of the regulations related to the award of the degree.

The results contained in this thesis have not been submitted either in part or in full to any other university or institute for the award of any degree or diploma to the best of my knowledge.

Dr. Shaunak Sen

Associate Professor,

Department of Electrical Engineering,

Indian Institute of Technology Delhi.

(Thesis Supervisor)

Acknowledgements

My life at IIT Delhi has been a true learning experience and I am grateful to many people for being a part of this.

First, I would like to express my heartiest gratitude for my supervisor, Prof. Shaunak Sen for his valuable guidance and support. He helped come up with the problems addressed in this thesis and always provided valuable feedback for the research work. He also gave me numerous important advises to develop skills such as writing and presenting. My research committee members, Prof. Indra Narayan Kar, Prof. S. Janardhanan and Prof. James Gomes have been immensely helpful giving insightful feedback over the course of last five years.

I acknowledge Atul Agarwal, Kushal Chakrabarti and Krishan Kumar Gola who did their M. Tech in Control and Automation group, IIT Delhi for collaborating in my work. I am grateful for having collaborated with Dr. Rishi Relan of Technical University of Denmark. I thank Venkat and Abhilash for critical reading of manuscripts and providing important feedback. I especially thank Joyjit, Nalin and all the former and current members of control and automation group for making my stay in IIT Delhi memorable. I thank Mr. Sunil and Virender of Control laboratory for their help in doing official works on so many occasions.

Finally, I thank my family for their eternal support. I always had the blessings of my late grandmother in whatever I did. My parents have always encouraged and supported me to pursue research, no words are enough to express my gratitude for them. My wife Thumree, being a researcher, always could relate to my worries and was still able to bring out the positive invariably. Without her love and encouragement my journey as a Ph.D. student would have been far more difficult.

Abhishek Dey

Abstract

Modeling and identification are integral in understanding any natural system and in designing new ones. In case of biomolecular systems, challenges arise due to their nonlinear, complex, and stochastic nature. Given these challenges, the general problems of estimating parameters from noisy measurements and obtaining approximate models in biomolecular systems are relatively unclear. Particularly, we answer three questions in this thesis relating to modeling and identification of biomolecular systems. Firstly, we address the problem of predicting bacterial growth response to dynamic environmental conditions that is important in designing robust synthetic circuits. Secondly, we address the challenge of parameter estimation in biomolecular systems from noisy data investigating the dependence of noise on system states and parameters. Thirdly, we address the problem of approximating nonlinear and complex biomolecular systems to useful linear models. To address these problems, we combine theoretical, computational, and experimental approaches. More specifically, to predict bacterial growth when temperature is varying dynamically, we estimate parameters in phenomenological nonlinear growth models at different fixed temperatures from experimental data and use these estimates for prediction. Next, we use the information of process noise dependence on states and parameters in the Langevin framework to adapt the standard Extended Kalman filtering algorithm by updating the process noise covariance at each time step. We find that the updated filter can achieve a balance between mean square estimation error and parameter convergence time and we discuss its optimality. Finally, we use the describing function-based technique to approximate nonlinear input-output maps in biomolecular systems using finite amplitude inputs. We find that the approximation error using describing function technique is lesser than standard linearisation and propose an analytical bound for this approximation error. The describing function-based approximation provides useful insight for the biomolecular systems investigated such as an account of the pure delay involved in a positive transcriptional feedback circuit, a bound of oscillation amplitude for biomolecular oscillators. Overall, the modeling and identification problems addressed in this thesis will aid in the analysis and design of biomolecular networks.

सार

मॉडलिंग और पहचान किसी भी प्राकृतिक प्रणाली को समझने और नए डिजाइन करने में अभिन्न हैं। बायोमोलेक्युलर सिस्टम के मामले में, उनके अरेखीय, जटिल और स्टोचैस्टिक प्रकृति के कारण चुनौतियां पैदा होती हैं। इन चुनौतियों को देखते हुए, नॉइस डेटा से पैरामीटर्स का अनुमान लगाने और बायोमोलेक्युलर सिस्टम में अनुमानित मॉडल प्राप्त करने की सामान्य समस्याएं अपेक्षाकृत अस्पष्ट हैं। विशेष रूप से, हम इस थीसिस में तीन सवालों के जवाब मॉडलिंग और बायोमोलेक्युलर सिस्टम की पहचान से संबंधित हैं। सबसे पहले, हम गतिशील पर्यावरणीय परिस्थितियों के लिए बैक्टीरिया के विकास की प्रतिक्रिया की भविष्यवाणी की समस्या को संबोधित करते हैं जो मजबूत सिंथेटिक सर्किट को डिजाइन करने में महत्वपूर्ण है। दूसरे, हम सिस्टम स्टेट्स और पैरामीटर्स पर नॉइज की निर्भरता की जांच नॉइस डेटा से बायोमोलेक्युलर सिस्टम में पैरामीटर आकलन की चुनौती को संबोधित करते हैं। तीसरा, हम उपयोगी रैखिक मॉडल के लिए गैर-रेखीय और जटिल बायोमोलेक्युलर सिस्टम को अनुमानित करने की समस्या को संबोधित करते हैं। इन समस्याओं को दूर करने के लिए, हम सैद्धांतिक, कम्प्यूटेशनल और प्रयोगात्मक दृष्टिकोणों को जोड़ते हैं। अधिक विशेष रूप से, बैक्टीरिया की वृद्धि की भविष्यवाणी करने के लिए जब तापमान गतिशील रूप से भिन्न हो रहा है, हम प्रायोगिक डेटा से अलग-अलग निश्चित तापमान पर घटनात्मक अरेखीय विकास मॉडल में मापदंडों का अनुमान लगाते हैं और भविष्यवाणी के लिए इन अनुमानों का उपयोग करते हैं। इसके बाद, हम हर समय कदम पर प्रक्रिया नॉइज सहसंयोजक को अद्यतन करके मानक विस्तारित कलामन फ़िल्टरिंग एल्गोरिदम को अनुकूलित करने के लिए लैंगविन फ्रेमवर्क में स्टेट्स और पैरामीटर्स पर नॉइज की निर्भरता की जानकारी का उपयोग करते हैं। हम पाते हैं कि अपडेट किया गया फ़िल्टर माध्य वर्ग अनुमान त्रुटि और पैरामीटर अभिसरण समय के बीच एक संतुलन प्राप्त कर सकता है और हम इसकी इष्टतमता पर चर्चा करते हैं। अंत में, हम परिमित आयाम इनपुट का उपयोग करते हुए बायोमोलेक्युलर सिस्टम में नॉनलाइनियर इनपुट-आउटपुट मैप्स के अनुमानित डिस्क्रिबिंग फंक्शन-आधारित तकनीक का उपयोग करते हैं। हम पाते हैं कि डिस्क्रिबिंग फंक्शन तकनीक का वर्णन करते हुए सन्निकटन त्रुटि मानक रैखिककरण की तुलना में कम है और इस सन्निकटन त्रुटि के लिए एक विश्लेषणात्मक बाउंड प्रस्तावित करता है। डिस्क्रिबिंग फंक्शन-आधारित सन्निकटन जैव-आणविक प्रणालियों के लिए उपयोगी अंतर्दृष्टि प्रदान करता है, जिसकी जांच एक सकारात्मक ट्रांसक्रिप्शनल फीडबैक सर्किट में शामिल शुद्ध विलंब के एक खाते के रूप में की जाती है, जो बायोमोलेक्यूलर ऑसिलेटर्स के लिए दोलन आयाम का एक समूह है। कुल मिलाकर, इस थीसिस में संबोधित मॉडलिंग और पहचान की समस्याएं बायोमोलेक्यूलर नेटवर्क के विश्लेषण और डिजाइन में सहायता करेगी।

Contents

Acknowledgements	iii
Abstract	v
List of Figures	ix
List of Tables	xvii
1 Introduction	1
1.1 Literature Review	2
1.2 Motivation	4
1.3 Problem Statement	4
1.4 Organization of this Thesis	5
1.5 Contribution of this Thesis	7
2 Dependence of Bacterial Growth on Temperature Changes	9
2.1 Background	10
2.2 Motivation	11
2.3 Problem Statement	11
2.4 Growth Models	12
2.5 Method	13
2.6 Estimation of Growth Parameters and their Temperature Dependence	14
2.7 Prediction of Bacterial Growth for Time-varying Temperature	16
2.8 Discussion	18

3	A Kalman Filter approach for biomolecular systems	21
3.1	Background	22
3.2	Motivation	23
3.3	Problem Statement	23
3.4	Mathematical Framework	23
3.5	Parameter Estimation for Biomolecular Systems	28
3.6	Performance Of Filter	34
3.7	Discussion	38
4	Describing function-based approximations of biomolecular systems	41
4.1	Background	41
4.2	Motivation	42
4.3	Problem Statement	43
4.4	Computation of Describing Function Approximation	43
4.4.1	Describing Function Method	43
4.4.2	Approximation for Other Realistic Inputs	56
4.5	Analytical Error Bound	58
4.6	Error Estimates in Limit Cycle Analysis	62
4.7	Nonlinear Distortion Analysis	67
4.8	Discussion	71
5	Conclusion	73
5.1	Summary of Contributions	73
5.2	Relevance of Work	74
5.3	Future work	75
	Bibliography	76
	Appendix	90

List of Figures

2.1	Prediction of bacterial growth profile with dynamic temperature variations. . . .	12
2.2	Bacterial growth for <i>E. coli</i> MG1655 for different temperatures. a. Bacterial growth at $29^{\circ}C$ in five different wells repeated for three days. b. Mean bacterial growth with daily variation in the temperature range from $29^{\circ} - 37^{\circ}C$	14
2.3	Different models for bacterial growth fitted to data. a. Comparison of different growth models fitted to experimental bacterial growth data at $29^{\circ}C$. b. Estimated maximum specific growth rate for different models at individual temperatures. c. Estimated carrying capacity for different models at individual temperatures. d. Other estimated model parameters at individual temperatures.	16
2.4	Predicted bacterial growth with and without using two sets of parameters of Verhulst - Pearl model for a step change in temperature. Blue line represents experimental growth with daily variation, red line signifies prediction with model parameters corresponding to lower temperature only and green line represents prediction using model parameters corresponding to both lower and upper temperature before and after application of step for a. step change in temperature from $29^{\circ} - 33^{\circ}C$ at 120 minute. b. step change in temperature from $29^{\circ} - 33^{\circ}C$ at 240 minute. c. Absolute error in prediction when step is applied at 120 minute. d. Absolute error when step is applied at 240 minute.	17

2.5	Predicted bacterial growth for dynamic variation in temperature. a. Experimental and predicted bacterial growth using different models when a step change from $29^\circ - 33^\circ C$ in temperature is applied at 120 minutes. b. Experimental and predicted bacterial growth using different models when a step change from $33^\circ - 37^\circ C$ in temperature is applied at 120 minutes. c. Experimental and predicted bacterial growth using different models when a step change from $30^\circ - 37^\circ C$ in temperature is applied at 120 minutes. d. Experimental and predicted bacterial growth using different models when temperature is varied in a staircase fashion from $29^\circ - 31^\circ - 33^\circ - 35^\circ C$ at 90, 150 and 210 minutes respectively. . .	19
3.1	State estimation for Example 1. Comparison of a. state estimates (inset shows zoomed in version), b. mean square estimation error for the three cases of process noise covariance, c. corresponding error covariances.	27
3.2	Effect of Q_2 . Mean estimate of last 100 time points along with standard deviation for fixed Q_1 , as well as updated $\hat{Q}_{1k}$ with an ensemble size $N = 10$ is plotted, the dashed black line shows the value of the unknown parameter used in simulation to generate data for a. $A_T = 10$, b. $A_T = 100$ and c. $A_T = 1000$	29
3.3	Estimated parameter K_2 for signaling system. Dashed black line shows true value of the parameter. Mean estimate of last 100 time points along with standard deviation for fixed Q_1 (blue line) as well as updated $\hat{Q}_{1k}$ (red dot) with an ensemble size $N = 10$ is plotted for a. $A_T = 10$, b. $A_T = 100$, and, c. $A_T = 1000$. $Q_2 = 10^{-7}$ in all three cases.	30
3.4	Estimation of parameter K in repressilator model. Black solid and dashed horizontal lines represent true parameter value and $\pm 5\%$ band of true parameter value. For each choice of process noise covariance, the filter algorithm is run 10 times. a. Parameter estimate using updated $Q_{i,k}$. b. Parameter estimate using fixed Q of the structure $Q_1 I_{6 \times 6}$	31

- 3.5 Estimated parameter α_s and K_d of *E. coli* heat shock response model. Blue lines represent fixed process noise covariance, red line represents updated process noise covariance and dashed black line represents true value of the parameter. a. Estimate of α_s , b. estimate of K_d when initial guess of parameters are set as 10. c. Estimate of α_s , d. Estimate of K_d when initial guess of parameters are set as 70. 32
- 3.6 Estimation of parameters from experimental data of negative transcriptional feedback circuit. Blue lines represent estimate using different fixed values of process noise covariances for, a. α and b. K . Red lines represent estimate using updated process noise covariance for, c. α and d. K . The value of measurement noise covariance, R is taken as 0.1 in all the cases. 35
- 3.7 Whiteness of the innovation sequence for Example 2 using autocorrelation test. The blue lines are mean estimated autocorrelation along with standard deviation and the black lines denote 95% confidence limit for subfigure a to f. Estimated autocorrelation of innovation sequence for different values of time-shift, a. for $A_T = 10$, using updated Q_{1k} . b. for $A_T = 100$, using updated Q_{1k} . c. for $A_T = 1000$, using updated Q_{1k} . d. for $A_T = 10$ using fixed $Q_1 = 1$. e. for $A_T = 100$ using fixed $Q_1 = 1$. f. for $A_T = 1000$ using fixed $Q_1 = 1$ 36
- 3.8 Whiteness of the innovation sequence for Example 2 using Box-Ljung-Pierce test. Red dotted line is the critical value and the blue markers represent the Q-statistics for subfigure a to f. a. $A_T = 10$, using updated Q_{1k} . b. $A_T = 100$, using updated Q_{1k} . c. $A_T = 1000$, using updated Q_{1k} . d. $A_T = 10$ using fixed $Q_1 = 1$. e. $A_T = 100$ using fixed $Q_1 = 1$. f. $A_T = 1000$ using fixed $Q_1 = 1$ 37
- 3.9 Trade-off between convergence time and MSE for Example -2, $A_T = 10$ for fixed Q_1 (blue line) as well as updated $\hat{Q}_{1k}$ (red dot). Convergence time is considered to be the time taken to reach $\pm 5\%$ band of true parameter value. Measurement noise (v_k) is an uniform integer random variable a. in $[-5, 5]$. b. in $[-1, 1]$ 38
- 3.10 Comparison of three cases of updated process noise covariance for Example 1. a. state estimates, b. corresponding mean square errors in estimation. 39

- 4.1 Approximation of the biomolecular system with saturating input-output map and corresponding error estimate. a. Black curve represents steady-state input-output response for parameters $A_T = 100nM$, $k_- = 100/hr$, $k_{+0} = k_-$, $b = k_{+0}/2$. Blue dot denotes the operating point for standard linearization. Schematic of the system is shown in the inset. b. Red triangle shaped, blue pentagon shaped and green circle shaped markers show the magnitude plots of analytical, computational describing function-based approximations and standard linearization respectively using same parameter set. c. Phase plots corresponding to b. d. Corresponding error plots. 45
- 4.2 Approximation of the biomolecular covalent modification system and error estimates. a. Solid black curve shows the steady-state normalized input-output response for parameters $A_T = 200nM$, $E_{1T} = E_{2T} = 20nM$, $d_1 = d_2 = 1/hr$, $k_2 = 1/hr$, $k_{10} = k_2$, $b = k_{10}/2$ and $a_1 = a_2 = 10^{-2}/hr$, which corresponds to lower sensitivity regime. Dashed black curve shows the steady-state normalized input-output map which corresponds to higher sensitivity regime, with $a_1 = a_2 = 1/hr$, other parameters being constant. Blue dot denotes the steady state operating point while computing standard linearization. b. Solid Blue and red curves correspond to the magnitude plots of computational describing function-based approximation and standard linearization of lower sensitivity regime respectively. Inset shows corresponding phase plots. c. Dashed Blue and red curves show magnitude and phase plots for both approximations in higher sensitivity case. d. Corresponding error in both approximations for both sensitivity regimes. 49

- 4.3 Describing function approximation and error estimate for biomolecular system with hysteretic input-output map. a. Solid red curve shows the steady-state normalized input-output response for parameters $\alpha_0 = 1/15$, $\alpha = 5nM/hr$, $n = 2$, $\gamma = 1/hr$, $K = 1nM$, $k_- = 100/hr$, corresponding to a hysteretic regime. Solid blue curve represents the same when $\alpha = 1$ with other parameters being same. Solid black vertical line denotes the upper limit of the area where describing function approximation is done for higher input amplitude, $k_{+0} = b = 375$. Dashed black vertical lines represents the upper and lower limit of the area where describing function approximation is done for lower input amplitude, $k_{+0} = 375, b = 300$. Schematic of the system is shown in inset. b. Solid Red and dashed red curves represent the magnitude plot of computational describing function-based approximation for hysteretic response with higher and lower input amplitude respectively and inset shows phase plot. c. Blue solid and dashed line shows the magnitude and phase plots for non-hysteretic regime with higher and lower input amplitude respectively. d. Corresponding error associated with the approximation in both regime for both higher and lower input amplitude. 51
- 4.4 Approximation of the metabolic gene regulatory network of *Saccharomyces cerevisiae* and error estimate. a. Solid red and dashed red lines are the magnitude plots for describing function and standard linearization at steady state respectively for the parameters for K699 strain (Bennett et al., 2008). Solid blue and dashed blue lines are the same for the parameters corresponding to YPH499 strain. $[glu]_{e0} = [gal]_e = 1000$, $b = 500$. b. Phase plots corresponding to a. c. Relative error plots for describing function approximation. d. Relative error plots for standard linearization approximation. 55

-
- 4.5 Approximation of a cascade network. a. Schematic of cascade network. b. Solid black, dashed black and dash-dotted black curves are steady state normalized input output maps at each stage for parameters $A_T = B_T = C_T = 100nM, k_- = 100/hr, k_{+0} = k_-, b = k_{+0}/2$. c. Solid black, dashed black and dash-dotted black lines are the magnitude plots for describing function for stage 1, 2 and 3 respectively. Inset shows corresponding phase plots. d. Solid black, dashed black and dash-dotted black lines are the magnitude plots for the describing function approximation of the input-output map from k_+ to C_1 , multiplication of describing function for each stage and multiplication of standard linearization for each stage respectively. Corresponding phase plots are shown in the inset. e. Solid black, dashed black and dash-dotted black lines are actual output (C_1), output when each stage is approximated with describing function and output when each stage is approximated by standard linearization respectively for input sinusoid with frequency = $0.2848cycles/hr$ 57
- 4.6 Application of describing function technique in estimating output to square wave input for the system in Example 1 for $\omega/2\pi = 20.0923$ cycles/hour. Square wave input is decomposed upto 5th harmonic and each harmonic is taken as a sinusoidal input for the use of describing function. Output to each harmonic is added to get the approximate output and the results are compared. 57
- 4.7 Error in describing function approximation and corresponding error bounds. a. Blue solid line and red dotted line represents mean squared error in describing function approximation and corresponding error bound respectively for static saturation, when ω is varied. The same is shown in dB scale in the inset. b. Similar plots for static hysteresis case. c. For the input-output map in Example 1, solid blue line represents the mean square error in the describing function approximation. Red dashed line represents the theoretical error bound when total variation is taken as $A_T/2$. Dashed blue line represents the bound when total variation is calculated computationally. 61

- 4.8 Limit cycle analysis of Van der Pol oscillator using describing function technique. a. Block diagram of Van der Pol oscillator. b. Solid blue line represents nominal describing function approximation whereas upper and lower limit for the approximation is represented by cyan and red line respectively. c. Black line represents the actual solution for Van der Pol oscillator. Blue, red and cyan line corresponds to the solution considering nominal, upper and lower limit of describing function approximation for $\mu = 0.4$. d. The same is repeated for $\mu = 1$. 63
- 4.9 Limit cycle analysis of repressilator using describing function technique. a. Block diagram of repressilator. b. Solid blue line represents nominal describing function approximation whereas upper and lower limit for the approximation is represented by solid red and dashed red line respectively for parameter values $\alpha = 0.4995 \text{ nM/s}$, $\alpha_0 = 5 \times 10^{-4} \text{ nM/s}$, $\gamma_m = \log 2/120 \text{ s}^{-1}$, $\gamma_p = \log 2/600 \text{ s}^{-1}$, $\beta = \log 2/6 \text{ s}^{-1}$, $K = 40 \text{ nM}$, $n = 2$. c. Black, blue and red lines represent the actual solution, nominal describing function solution and describing function solution considering the upper error limit for repressilator when $\gamma_p = 20 \log 2/600$ and other parameters are same as b. d. The same is repeated for $\gamma_p = 14 \log 2/600$. 66
- 4.10 Frequency spectrum of output for two-state signaling system for parameters $A_T = 100 \text{ nM}$, $K_- = 100/\text{hr}$ and different values of RMS of multisine. Black, blue, green and magenta dots represent the noise, excited frequencies, even harmonic and odd harmonic distortion respectively. 69
- 4.11 Nonlinear distortion analysis for G-K switch. Left column is for parameter $a_1 = a_2 = 1/\text{nM} - \text{hr}$ and right column for $a_1 = a_2 = 100/\text{nM} - \text{hr}$. (a) and (b) show normalized input and normalized output map. (c), (d) and (e), (f) are frequency spectrum of A^* for RMS = 1 and 30 respectively. The value of other parameters is: $A_T = 200 \text{ nM}$, $E_{1T} = E_{2T} = 20 \text{ nM}$, $d_1 = d_2 = 100/\text{hr}$ and $k_2 = 100/\text{hr}$. Black, blue, green and magenta dots represent the noise, excited frequencies, even harmonic and odd harmonic distortion respectively. 70

List of Tables

2.1	Mean of coefficient of determination (R^2)/ mean square error(MSE) between experimental data and different growth models at individual temperatures. . . .	15
2.2	Mean of coefficient of determination (R^2)/ mean square error(MSE) between experimental and predicted bacterial growth profile for different growth models subjected to dynamic temperature variation.	19
3.1	Percentage of estimated autocorrelation values outside 95% confidence limits for biomolecular systems.	36
3.2	Mean square estimation errors for biomolecular systems.	37