

The Application of Nonlinear System Identification in the Field of Synthetic Biology

A thesis submitted to the University of Sheffield for the degree of Doctor of
Philosophy

Kirubhakaran Krishnanathan

Department of Automatic Control and Systems Engineering

March 2014

To my family

Acknowledgements

I would like to express my sincere gratitude to my supervisors Professor Visakan Kadiramanathan and Professor Stephen A. Billings for providing me with the excellent opportunity to work on this exciting multidisciplinary project. Their support and encouragement as well as their patience and trust were essential for completing this thesis. I would like to thank Prof. Visakan for all the invaluable advice and project related ideas, and particularly for guiding me in the right direction. Equally, I would like to thank Prof. Steve for letting me greatly benefit from his great expertise and knowledge on nonlinear system identification, and for being supportive since my undergraduate study.

I would like to thank Professor Phillip Wright for providing me laboratory space and equipments, and encouraging me to perform and design experiments.

I would like to thank Dr. Sean Anderson for technical discussions, comments and his encouragement which have made a huge impact on my progress. I am very much indebted to him.

I am particularly grateful to Stephen Jaffe who not only helped me to get started with experiments but also for all the light hearted chats that enriched the G4B work atmosphere.

Many thanks to everyone, past and present, in room 316 for providing such a great place to work in. Special thanks goes to Dr. Andrew Hills for all the help in computer related problems.

A big thank you to all my friends who have made Sheffield a special place.

Finally, I would like to show my appreciation to my family who have supported me throughout my entire university studies and I am always grateful for all that they have given me. Without them, none of this would have been possible.

Abstract

The field of synthetic biology has progressed from early concept, to initial demonstrations of simple genetic parts, and more recently to biological systems composed of functional modules that perform useful and specified tasks. Globally, there is an expectation that synthetic biology will deliver solutions to challenges, for instance, in healthcare, food security, and energy production. A key challenge in synthetic biology is to develop effective methodologies for characterisation of modular genetic parts in a form suitable for dynamic analysis and design. Dynamic analysis will enable the design of genetic parts to achieve robust and extensive functionalities, unlike the more commonly applied static analysis.

In this thesis, improvements and new designs of both experimentation and modelling methods are presented, which were used for the quantitative analysis of transcriptional regulatory genetic parts and the development of mathematical models to aid predictive model-based design of higher-order genetic parts, in a top-down design approach.

A data-driven nonlinear dynamic modelling framework is proposed to identify dynamic models of genetic parts. The identified models are shown to have compact representation and achieve rapid, accurate prediction of experimental data. The identification framework was extended by incorporating a computational Bayesian approach, to estimate the uncertainty of model parameters. The novel identification framework was used to capture the cell population heterogeneity observed in experimental data of the systems.

To investigate if a reporter cascade has an influential effect on the dynamics of the system to which it has been linked to, the identification framework was used to characterise dynamics of two transcriptional regulatory systems - the same functional module but different reporter cascades. For the first time this provided evidence that the reporter cascades do have an influential effect on the dynamics of the systems. Generalised frequency response functions obtained from the identified dynamic models provided an alternative tool for dynamic characterisation of genetic parts which could be used for design purposes. In addition, characterising only the functional module - BBa_F2620 relative to a reporter cascade was found to be unachievable using the implemented experimentation. However, with the identification and analysis tools used, the commonality of the systems under investigation is retrieved and adequately characterised.

Contents

Nomenclature	xv
1 Introduction	1
1.1 Background	1
1.2 Motivation	3
1.3 Overview of the thesis	6
1.4 Research outputs arising from this thesis	7
1.5 Summary of contributions	8
2 Synthetic biology as an engineering problem	9
2.1 Introduction	9
2.2 An engineering problem with great complexity	10
2.2.1 Standardisation, decoupling and abstraction	13
2.2.2 Characterisation and model-based design	15
2.2.3 Variation in gene expression	16
2.3 Biochemical modelling	16
2.3.1 Cell growth models	16
2.3.2 Gene expression models	18
2.4 Summary	22
3 System identification and its literature	23
3.1 Introduction	23
3.2 Model structures	24
3.2.1 Linear black-box models	25
3.2.2 Nonlinear black-box models	26
3.2.3 Cascade models	28
3.3 Parameter estimation	30
3.3.1 Statistical approach	31
3.3.2 Non-statistical approach	33
3.4 Model structure detection	35

3.4.1	Linear regression	36
3.4.2	Nonlinear regression	38
3.5	Model selection	39
3.6	Model validation	40
3.7	Continuous-time system identification	40
3.7.1	Model estimation for continuous-time system identification	41
3.7.2	Signal derivative estimation using dCTM	42
3.8	Generalised frequency response functions	43
3.8.1	Probing method	44
3.9	Summary	45
4	Modelling a transcriptional regulation	46
4.1	Introduction	46
4.2	The BBa_T9002 system and its experimental data	48
4.3	Data pre-processing - signal derivative estimation	49
4.3.1	Kalman filter	50
4.3.2	Rauch-Tung-Striebel smoother	50
4.4	Identification of dynamic and static biochemical models	51
4.4.1	Derivation of dynamic and static biochemical models	51
4.4.2	Parameter estimation of enzymatic reaction scheme model and Hill equation	54
4.4.3	Results and discussion	56
4.5	Identification of data-driven dynamic model	58
4.5.1	System description and physical insights	59
4.5.2	CT-NARX model representation	59
4.5.3	CT-NARX model with static input nonlinearity	61
4.5.4	Parameter estimation and structure detection of CT-NARX model	62
4.5.5	Results and discussion	63
4.6	Summary and further discussion	68
5	A novel identification framework for continuous-time non-linear dynamic systems	71
5.1	Introduction	71
5.2	Parameter estimation by approximate Bayesian computation	74
5.2.1	Rejection sampling	75
5.2.2	Inefficiency of the basic ABC	75
5.3	Model definition and parameter estimation	76
5.3.1	Continuous-time nonlinear model representation	76

5.3.2	Parameter estimation by ABC-SMC	77
5.3.3	Implementing ABC-SMC for nonlinear continuous-time model	77
5.4	Nonlinear continuous-time model identification framework	79
5.4.1	One-stage model structure detection	79
5.4.2	Two-stage model structure detection	81
5.4.3	Derivative order model selection	83
5.5	Multi-core processing for fast ABC	84
5.6	Results	84
5.6.1	Case 1: parameter estimation of VDPO system	85
5.6.2	Case 2: model structure detection of VDPO system	87
5.6.3	Case 3: model structure detection and derivative order model selection of a test system with identifiability problem	90
5.7	Summary	91
6	System fabrication, experimental design and data acquisition	95
6.1	Introduction	95
6.2	Experimental protocols	97
6.2.1	Transformation	98
6.2.2	Preparation of M9 supplemented media	100
6.2.3	Colony screening	101
6.2.4	Diagnostic gel	101
6.2.5	DNA assembly protocols	102
6.2.6	Qiagen toolkits	103
6.3	Design and assembly of genetic parts	103
6.3.1	BBa_F2620 and BBa_T9002	105
6.3.2	BBa_J06702 and newly designed genetic part	105
6.4	Experimental setup and data acquisition procedures	110
6.5	Overview of experimentally obtained datasets	112
6.6	Summary	112
7	Interpretation of a gene reporter signal and key dynamic design properties	115
7.1	Introduction	115
7.2	Biological differences in the reporter cascades	119
7.3	Modelling BBa_T9002 and "F2620-RC2" systems	119
7.3.1	Experimental data	121
7.3.2	Model representation	121
7.3.3	Parameter estimation and structure detection of system model	126
7.4	Results and discussion	127

7.4.1	Cell growth properties	128
7.4.2	"Single-cell" protein expression properties	130
7.4.3	Model validation of a unified model with parameter uncertainty	136
7.5	Summary	137
8	Conclusions	139
8.1	Conclusions and summary	139
8.2	Future work	142
A	GFRFs computation of the "single-cell" model of BBa_T9002 system	144
B	GFRFs computation of the "single-cell" model of "F2620-RC2" system	146
	Acronyms	148
	Bibliography	150

List of Figures

1.1	The cost of synthesising genes and oligos (Carlson, 2009).	2
1.2	Design stages envisioned by the synthetic biology community. . . .	3
2.1	Transcription and translation steps in a) prokaryotic cells and b) eukaryotic cells (image obtained from openstax college free online book - biology). In this thesis, only bacterial cells are considered, which are prokaryotic cells. During transcription, RNA copy of a gene sequence is made, called the messenger RNA that is read by ribosomes, in order to translate it into a sequence of amino-acids during protein synthesis.	12
2.2	A snapshot of the catalogue in the registry of standard biological parts	14
2.3	The typical growth of number of cells in a microbial culture	17
2.4	Examples of CGD models. A. Models of the repressilator (left) (Elowitz and Leibler, 2000) and Lac operon (right) (Lestas et al., 2008), and B. A snapshot of the CySBML, which provides a platform to import SBML files (http://apps.cytoscape.org/apps/cysbml).	21
3.1	Cascade models a) Hammerstein model and b) Wiener model.	29
4.1	Pictorial description of the BBa_T9002 system with input and output of 3OC ₆ HSL and GFP expression respectively.	48
4.2	The GFP expression signal and its derivatives obtained from the RTS smoothing algorithm (black) in comparison to derivatives from numerically differencing the raw GFP expression signal (grey).	52

- 4.3 Biochemical model simulation for experiment 1: (A) The simulated model input signal $s(t)$, (B) Rate of change of GFP expression at the 150th minute (blue), the Hill equation prediction (green), and the ERS model prediction (red), (C) Comparison of GFP expression (blue) and the model prediction $p(t)$ (red) and (D) Rate of change of GFP expression (blue) and model prediction (red). Note that the response corresponding to the lowest input level $3OC_6HSL = 0$ in (A) has been omitted because of the log transformation, and the plots in (C) and (D) corresponds to responses to increasing input level from left to right - top to bottom. 57
- 4.4 System representation and physical insight: (A) The model structure representation of the data-driven model where the dynamic function corresponds to the CT-NARX model and (B) The individually normalised GFP expression of each response in experiment 1. Normalisation is with respect to the final GFP expressions to remove the static gain effects. 60
- 4.5 CT-NARX structure detection: (A) Mean squared prediction error (MSSE) for CT-NARX models with an MSSE < 5 and (B) Akaike and Bayesian information criteria (AIC and BIC respectively), optimal model with minimum AIC and BIC value is model structure 16. Note in both (A) and (B), the models are ordered by increasing complexity, i.e, number of model terms. 64
- 4.6 CT-NARX model simulation: (A) CT-NARX input signals, (B) Rate of change of GFP expression at the 150th minute (blue), the Hill equation prediction (green), the CT-NARX model prediction at observed input concentration (red stars), and the CT-NARX model prediction at interpolated input concentrations (red crosses), (C) Comparison of GFP expression (blue) and the CT-NARX prediction (red) and (D) Rate of change of GFP expression (blue) and CT-NARX model prediction (red). Note that the response corresponding to the lowest input level $3OC_6HSL = 0$ in (A) has been omitted because of the log transformation, and the plots in (C) and (D) corresponds to responses to increasing input level from left to right - top to bottom. 66

- 4.7 CT-NARX model identification across colonies and experimental data sets. (A-C) Estimates of CT-NARX model dynamic parameters c_1, c_2 and c_3 . (D) CT-NARX model mean sum of squared error (MSSE) for each of 6 different experimental data sets where sets are grouped by colony: colony 1 comprises experiments 1-3; colony 2 comprises experiments 4-6. 67
- 4.8 Static model of the input nonlinearity. The CT-NARX dynamic model input $\tilde{u}(t)$ was obtained from transforming $u_*(t) = \log_{10}(gu(t))$ through a static function $G(u_*(t))$, where $u(t)$ was the level of 3OC₆HSL. (A) Separate estimates of the static function $G(\cdot)$ (red) across 6 experimental data sets (blue) used for identification purpose and (B) Single estimate of the static function $G(\cdot)$ (red) using experimental datasets compared to the average of the experimental data curves in panel (A) (blue dots) 68
- 4.9 CT-NARX model prediction on validation data. (A) A single CT-NARX model with average parameter estimates was simulated (red) and compared to a reserved set of validation data (blue) and (B) The percentage prediction error variance from the averaged CT-NARX model using both estimation (blue) and validation (red) datasets. . . 69
- 5.1 Term selection via the cumulative density function. The cumulative density function for a parameter θ is constructed by the ABC-SMC estimation algorithm. The model term is rejected if zero lies between the limits corresponding to the 5% and 95% probability levels, *i.e.* $a \leq 0 \leq b$ 80
- 5.2 Computation time of 4th order Runge-Kutta simulation algorithm. 1000 random VDPO simulations were repetitively simulated using the *parfor* function with varying processing cores. 85
- 5.3 Parameter estimation of VDPO system using ABC-SMC (Algorithm 5.1). True parameters are shown as red stem plots and the black dotted lines indicates the prior. Estimated sample distributions are shown over 10 iterations of the ABC-SMC procedure on VDPO system with measurement noise for SNR of 20dB - (A) and 10dB - (B) measurement noise. Iteration 1 in grey while iteration 10 in black. . 86
- 5.4 Model structure detection using one-stage procedure (Algorithm 5.2: $L = 3$ and $N_s = 200$) for VDPO system. True model terms in red stem are correctly selected, while false model terms in black stem are correctly not selected. The black dotted and red solid vertical lines indicates the prior and quantile values respectively. . . . 88

- 5.5 Comparison of noise free output (blue), dCTM model (green) and one-stage ABC (red - the shaded region indicates uncertainty from the ABC parameter range). A: one-stage ABC model based on initial parameter estimation (full model set) and B: one-stage ABC model based on re-estimation of parameters (only selected model terms). 89
- 5.6 Model structure detection using two-stage procedure (Algorithm 5.3: $L = 3$ and $N_s = 200$) and model selection procedure (A-D and F-G shows results for SNR=10dB). A. Cha-Srihari measure of the 12 model terms whose variance > 1 (the correct model terms are indicated), B. The consequent BIC score of models when model terms are added in order of sensitivity (the correct model terms are indicated and the BIC score under 0 represents the model when no terms from pool 2 are added), C. The variance relative to zero of each model term in the system is shown, indicating their contribution to the dynamics (the bar under 0 quantifies that of the system's output), D. The derivative order model selection procedure of model order's $n_i = (2,3)$, E and F. Comparison of noise free output (blue), dCTM model (green) and two-stage ABC (red - shaded region indicates uncertainty from ABC parameter range): E - SNR = 20dB and F - SNR - 10dB, and G. The posterior distribution (histogram) of the the 12 model terms whose variance > 1 (the correct model terms are indicated). 92
- 6.1 Equipments and hyperladder 1 chart used during experimental protocols. A. Large gel kit used for DNA purification or gel extraction, B. Compact gel kit used for diagnostic gel, C. Ultraviolet imaging system, D. Ultraviolet illuminator, E. Electroporation device and F. Hyperladder 1 chart - BIOLINE. 104
- 6.2 The BBa_F2620 and BBa_T9002 constructs. A. The standard plasmid construct (red) of genetic parts in RSBP, where E, X, S and P stands for the restriction sites EcoRI, XbaI, SpeI and PstI respectively, and the antibiotic tag is usually ampicillin, B. The DNA samples obtained from maxiprep protocol is viewed under the ultraviolet imaging system, BBa_F2620 (1061 base pairs) - yellow circle and BBa_T9002 (1945 base pairs) - red circle, C. and D. Snapshots from FinchTV showing the DNA sequence traces of BBa_F2620 and BBa_T9002 respectively (top window showcases the sequence obtained using forward primer, while the bottom window showcases the sequence obtained using the reverse primer). 106

- 6.3 The design steps and construct of the new genetic part - "F2620-RC2" system. At top, the pictorial description of the BBa_J06702 genetic part is shown. Below, the design steps used in constructing the new genetic part is shown, where the restriction sites X - XbaI and S - SpeI have similar sequences thereby complimenting each other. 108
- 6.4 Gel imaging, sequencing traces of BBa_J06702 and TECAN GENios microplate reader. A. The DNA sample obtained from maxiprep protocol is viewed under the ultraviolet imaging system, BBa_J06702 (869 base pairs) - red circle, B. Snapshots from FinchTV showing the DNA sequence traces of BBa_J06702 (top window showcases the sequence obtained using forward primer while the bottom window showcases the sequence obtained using the reverse primer), C. The double digested DNA samples of BBa_J06702 (red) and BBa_F2620 (yellow), D. The gel fragments excised for carrying out gel extraction, E. The DNA sample obtained from miniprep protocol is viewed under the ultraviolet imaging system, "F2620-RC2" system (1930 base pairs) - red circle, F. The TECAN GENios microplate reader and G. A 96 well plate prototype. 109
- 6.5 Experimental data of BBa_T9002. A. and B. First row - BBa_T9002 colony 1, second row - BBa_T9002 colony 2, third row - BBa_T9002 colony 3, and fourth row - BBa_F2620 colony 1 (control). Response due to 0 M of 3OC₆HSL in light green while response due to 1e-4 M of 3OC₆HSL in dark green. 113
- 6.6 Experimental data of "F2620-RC2". A. and B. First row - "F2620-RC2" colony 1, second row - "F2620-RC2" colony 2, third row - "F2620-RC2" colony 3, and fourth row - BBa_F2620 colony 1 (control). Response due to 0 M of 3OC₆HSL in light red while response due to 1e-4 M of 3OC₆HSL in dark red. 114
- 7.1 Pictorial description of the investigation. The G blocks and H blocks represent functional modules and reporter cascades respectively. In this investigation G₁ - BBa_F2620, H₁ - BBa_E0240, H₂ - BBa_J06702 and G₂ - arbitrary functional module. A. and B. is implemented to achieve the characterisation of G₁ relative to H₁ and H₂ respectively, while C. demonstrates the assembly of a arbitrary higher-order genetic part. 116

- 7.2 System representation. The model structure representation of the system model - blue dashed box, where the dynamic function (data-driven) corresponds to the CT-NARX model, the cell growth model corresponds to the modified Lin's model and the "single-cell" model is represented in the red dashed box. 122
- 7.3 Mean model simulation for BBa_T9002 system. A. Comparison of growth response (blue) and the modified Lin's model prediction (green), B. Static model of the input nonlinearity, estimate of the static function $G(\cdot)$ (green) compared to the experimental data curves (blue), C. Comparison of "single-cell" GFP expression (blue) and "single-cell" model prediction (green) and D. Comparison of GFP expression (blue) and system model prediction (green). Note that the response corresponding to the input levels $3OC_6HSL = 0$ (in B) and $1e-4 M$ (all) have been omitted because of the log transformation and outlier observation respectively. 132
- 7.4 Mean model simulation for "F2620-RC2" system. A. Comparison of growth response (blue) and the modified Lin's model prediction (red), B. Static model of the input nonlinearity, estimate of the static function $G(\cdot)$ (red) compared to the experimental data curves (blue), C. Comparison of "single-cell" RFP expression (blue) and "single-cell" model prediction (red) and D. Comparison of RFP expression (blue) and system model prediction (red). Note that the response corresponding to the input levels $3OC_6HSL = 0$ (in B) and $1e-4 M$ (all) have been omitted because of the log transformation and outlier observation respectively. 133
- 7.5 Bode plots (both magnitude and phase) of the first-order GFRFs of both BBa_T9002 (green) and "F2620-RC2" (red) systems, using mean values of the parameters. 135
- 7.6 BBa_T9002 system model variation. Comparison of GFP expression used for identification purpose (dark green line), GFP expression of other experiments (blue) and identified system model (light green shaded region - indicates uncertainty from the ABC parameter range). The plots corresponds to responses to increasing input level from left to right - top to bottom. 137

-
- 7.7 "F2620-RC2" system model variation. Comparison of RFP expression used for identification purpose (dark red line), RFP expression of other experiments (blue) and identified system model (light red shaded region - indicates uncertainty from the ABC parameter range). The plots corresponds to responses to increasing input level from left to right - top to bottom. 138

List of Tables

3.1	Some common linear black-box models	26
3.2	Interpretation of the Bayes factor	40
4.1	Mean and variability in CT-NARX model parameters across colonies	67
5.1	Estimated parameters of the VDPO system.	86
5.2	Identified models of the VDPO system. ABC1 refers to the one-stage ABC identification method.	87
5.3	Identified models from the test system in case 3. ABC1 refers to the one-stage ABC identification method, and ABC2 refers to the two-stage ABC identification method.	93
7.1	Differences in the reporter cascades. The properties with superscript indicated as <i>i</i> or <i>ii</i> are found from RSBP or (Shaner et al., 2005) respectively.	120
7.2	Parameters (mean values) of the cell-growth models for both BBa_T9002 and "F2620-RC2" systems.	129
7.3	Parameters (mean values) of the "single-cell" models (CT-NARX) for both BBa_T9002 and "F2620-RC2" systems.	130

Nomenclature

A list of the variables and notation used in this thesis is defined below. The definitions and conventions set here will be observed throughout unless otherwise stated. For a list of acronyms, please consult page 148.

θ	parameter vector
λ	standard deviation of noise signal
\mathbb{R}	real space
\mathbf{e}	noise vector
\mathbf{u}	input vector
\mathbf{y}	output vector
\mathbf{z}	undisturbed output vector
\mathcal{N}	normal distribution
n	Hill coefficient
$p(t)$	product concentration at time t
$s(t)$	substrate concentration at time t
$x(t)$	cell concentration at time t
$\mu(t)$	growth rate at time t
\top	matrix transpose
I	identity matrix
k	discrete time
L	number of iterations in the ABC framework

N_θ	maximum number of parameters
n_i	derivative order
N_s	number of parameter samples in the ABC framework
N_y	number of data samples
q	polynomial order
T	sampling time
t	time
Φ	regression matrix

Chapter 1

Introduction

1.1 Background

The multidisciplinary field - "synthetic biology", has been mentioned several times in news, documentaries and conferences related to global issues. The field has been growing in popularity and demand, and is now expected to deliver solutions to global challenges, for instance, in healthcare (Khalil and Collins, 2010, Lu and Collins, 2009, Weber et al., 2008), food security (Stocker et al., 2003, van der Meer and Belkin, 2010) and energy production (Alper and Stephanopoulos, 2009, Atsumi et al., 2008, Keasling and Chou, 2008). The food industry in particular, has embraced this new field and the technology it brings forward, as there are ever growing numbers of genetically modified agricultural products. However, the transparency and reliability of this new technology have been questioned by several, preventing genetically modified products in most sectors from becoming marketable (Engel et al., 1995).

So what is synthetic biology and what makes it so promising? Synthetic biology can be defined as the engineering of biology, which helps incorporate new functionalities in living cells that do not naturally occur in nature. The basic ideology of design and build in synthetic biology is to identify the target protein and construct the genetic network pathway required to produce the required target protein. This is a challenging task, as it comes with high complexity and uncertainty. However, it is still regarded as the key to solving global challenges. It attracts the minds of researchers from different disciplines like chemistry, physics, microbiology, mathematics and engineering (Endy, 2005).

Synthetic biology research is conducted in several laboratories worldwide. An

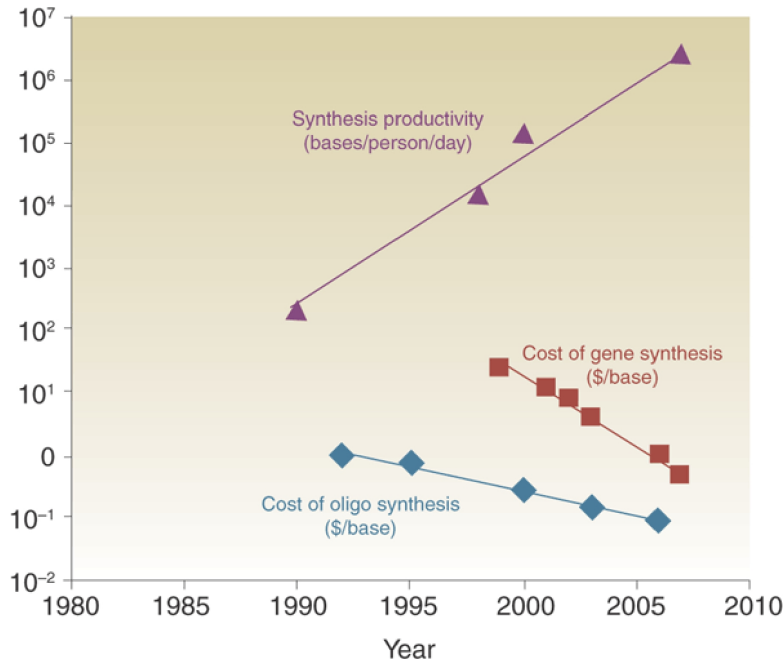


Figure 1.1: The cost of synthesising genes and oligos (Carlson, 2009).

obvious reason for synthetic biology's growth is the potential the field holds in transforming many sectors. However, another key reason behind the rapid growth is the reduced cost and time over the years, of synthesising genes and oligos (Figure 1.1). As reported in Carlson (2009),

"the number of bases a single individual can synthesize in a day using commercial instrument has increased by five orders of magnitude, whereas the per base cost of synthetic genes has dropped by nearly three orders of magnitude".

Engineering is gradually emerging to play a key role in synthetic biology (Endy, 2005). The general principles practiced in engineering are transferred to synthetic biology, especially in the design stages, such as drawing in-depth knowledge (first principles) which permits computer-based iterative design, implementation of decoupling and abstraction which enables individuals of different expertise to work independently in a hierarchical manner and finally, push the motive of standards in genetic parts to guarantee compatibility in design.

In this thesis a systems and control engineering approach is introduced to the field of synthetic biology. Nonlinear system identification, a commonly practiced approach in systems and control engineering, forms a backbone in adopting a model-based characterisation of genetic parts. It is demonstrated on application

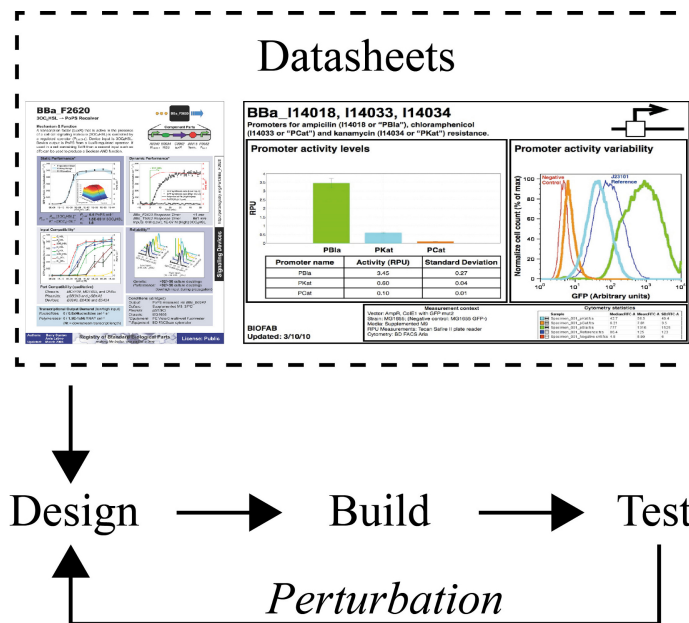


Figure 1.2: Design stages envisioned by the synthetic biology community.

that reveals the challenges faced in designing genetic parts.

1.2 Motivation

The approach to the design of genetic parts can be categorised into 2 types, bottom-up design and top-down design. Designing a higher-order genetic part by building from the very basic genetic network is termed the bottom-up design, whereas building using off-the-shelf fabricated genetic parts to obtain the required functionality is referred to as top-down design. Regardless of the design approach, the design stages envisioned by the synthetic biology community (Figure 1.2) are mostly practiced (Chandran et al., 2009).

Datasheets obtained through characterisation (Arkin, 2008, Canton et al., 2008), serve as catalogues to the characterised genetic parts, which summarises the necessary information needed for design. The information provided helps researchers build higher-order genetic parts, which are eventually tested on completion. The design stages: design, build and test are repeated until design specifications are met. However, there are a number of obstacles to overcome before realising such effective design in practice. In particular, challenges in characterisation are directly linked to our ability to derive useful models of genetic parts because the concept of implementing model-based design to assist the design stages of genetic parts is

central to the engineering ethos.

Model-based design helps in reducing trial-and-error design practice, saving time and cost (Ellis et al., 2009). Predictive and optimal characteristics of model-based design, helps in achieving this. Assuming a higher-order genetic part is set out to be designed, a knowledge base of specifically selected genetic networks and already fabricated genetic parts are collected. This permits computer-based extensive simulation to formulate if, the required functionality of the higher-order genetic part can be obtained. The predictive nature of the model-based design underpins the methodology. Once the higher-order genetic part is assembled, certain features of the higher-order genetic part or environmental perturbations are tuned repeatedly and simultaneously in both in-silico and in-vivo prototypes. This allows for robust design which showcases the desirable attributes of model-based design.

Existing models in the synthetic biology literature, have provided a shortcoming in the effectiveness of model-based design of genetic parts. Static functions such as Michealis-Menten, Hill equation *etc.*, which are simple but limited, retain a fixed model structure regardless of the complexity of the system under investigation, are mostly used. Dynamic models in the form of ordinary differential equations (ODEs) and stochastic differential equations (SDEs) have also been explored, which take predefined model structures that grow in size as the gene network topology increases. These models encounter large model complexity, constrained parameters and poor prediction accuracy to experimental data.

In this thesis a methodology is proposed, where a data-driven nonlinear dynamic modelling framework is developed to derive time-domain models of genetic parts, which are subsequently transformed to frequency-domain models. These models should be specified in datasheets as extensions, with the purpose of aiding the design and synthesis of higher-order genetic parts. The time-domain models used in this thesis, are a class of the continuous-time (CT) nonlinear autoregressive moving average model with exogenous input (NARMAX) (Billings, 2013), which will provide a solution for overcoming the typical problems of models for genetic parts that are overly complex, unwieldy, and of unknown structure. The generalised frequency response function (GFRF) is used to represent time-domain models in frequency-domain (Billings, 2013), enabling spectral analysis that will guide the identification and interpretation of dominant system characteristics of the genetic parts.

The time-domain models are derived to capture the dynamic input-output characterisation of genetic parts, opposed to modelling and characterising each of the subcomponents and from these predict the whole dynamics of the genetic parts which is equally consistent (Arkin, 2008). The time-domain models derived in this thesis are predictive but do not give a biochemical interpretation. However, input-output characterisation allows for quantitative analysis to be implemented which are tractable, thereby making re-design of genetic parts less challenging.

In order to achieve its aim (the methodology), this thesis accomplishes the following objectives:

- the implementation of a data-driven nonlinear dynamic modelling framework for the dynamic characterisation of genetic parts. The BBa_T9002 system was used as a case study, a simple quorum sensing composite genetic part. Extending its results in static input-output model characterisation (Canton et al., 2008) to a data-driven nonlinear dynamic model, a model representation that is compact and performs high accurate prediction is achieved.
- the development of a computational Bayesian identification framework to address the lack of qualitative study of variations observed in different cell population of BBa_T9002 system, which are due to population heterogeneity and gene noise. The developed identification framework provides the uncertainty in model parameters by constructing distributions, which captures the variation in the experimental data.
- the further experimentation of BBa_T9002 and "F2620-RC2" systems. The cell growth and protein expression measurements of both systems were collected during lag, exponential, stationary and decay phase of microbial growth. The implementation of the above proposed methodology towards the newly collected experimental data, allowed for a more robust dynamic characterisation of transcriptional regulatory genetic parts. It also paved way for great insights about the effects of using different ribosome binding sites and fluorescence proteins for the characterisation of transcriptional regulatory genetic parts.

1.3 Overview of the thesis

The thesis is structured into 8 chapters, covering: literature on the subjects of synthetic biology and system identification; development of new experimentation and mathematical methods; and their application to provide new insights into model-based predictive design of genetic parts in biological systems.

- Chapter 2 provides an overview of the similarities between synthetic biology and engineering. The inherent complexity in designing genetic parts in biological systems is discussed, and the need of implementing model-based design to assist design stages of biofabrication. Models of cell growth and gene expression are reviewed along with their shortcomings.
- Chapter 3 provides an in-depth review of techniques for system identification: parameter estimation, model structure detection and model selection. The need to identify models in continuous-time is emphasised, as key analytical design properties need to be shared with and understood by the biologists. In the last section, generalised frequency response function used for the spectral analysis of nonlinear black-box models are introduced.
- Chapter 4 proposes a data-driven framework to identify a nonlinear black-box model for dynamic characterisation of genetic parts in biological systems. An enzymatic reaction scheme model and continuous-time nonlinear autoregressive model with exogenous input are identified to characterise a transcriptional regulatory genetic part - BBa_T9002 system. The superior performance of the nonlinear black-box model is demonstrated in real experimental data. It was established that: (i) additional experimental data was required to robustly characterise the genetic part, where the experimental data is required to capture the dynamics (both cell growth and protein expression measurements) of the system through all phases of the microbial growth and (ii) a principled method is required to capture the cell population heterogeneity observed in biological systems.
- Chapter 5 introduces a computational Bayesian identification framework for nonlinear continuous-time systems that utilises a simulation approach. The main contribution of this algorithm to the suite of methodology available for continuous-time nonlinear system identification is the signal derivative free approach and the estimation of the model parameter uncertainty by constructing a distribution. The identification algorithm uses the approximate Bayesian computation - sequential Monte Carlo method, which generate parameter distributions that drive term selection by significance testing.

- Chapter 6 provides the experimental protocols needed to carry out the bio-fabrication process of the systems - BBa_T9002 and "F2620-RC2". The experimental setup and procedures used in acquiring the required experimental data are outlined. The collected experimental data is presented, which: (i) consist of both cell growth and protein expression measurements of the systems for longer time period and (ii) will aid in the investigation of whether a reporter cascade have an influential effect on the dynamics of the whole system it has been linked to.
- Chapter 7 investigates if reporter cascades are appropriate for characterisation. It was concluded in the chapter that, the reporter cascades do have an influence on the "relative" dynamics of both systems and characterising only the functional module - BBa_F2620 relative to a reporter cascade as an unachievable task using the implemented investigation in this thesis. However, with the identification and analysis tools used, the commonality of the systems under investigation (same functional module linked to two different reporter cascades) is retrieved and adequately characterised. It was also concluded that, more experiments with systems made of the functional module - BBa_F2620 and different reporter cascades have to be conducted, to deduce if the unique representation in time-domain of the "single-cell" protein expression dynamics obtained in the chapter can be used to represent a system made of the functional module - BBa_F2620 and an arbitrary reporter cascade to aid model-based design. Therefore, leaving the debate - appropriateness of reporter cascades for the use of characterisation of genetic parts, open to the synthetic biology community. The results presented hint at the possibility that dynamic characterisation with predictive ability can lead to new design tools in synthesising functional bioparts and devices.
- Chapter 8 concludes the work done in this thesis, and provide suggestions for future directions of research.

1.4 Research outputs arising from this thesis

Material from this thesis has formed the basis for one published paper and a book sub-chapter from Chapter 4 of this thesis,

- K. Krishnanathan, S. Anderson, S.A. Billings, and V. Kadiramanathan. A data-driven framework for identifying nonlinear dynamic models of genetic parts. *ACS synthetic biology*, 1(8):375384, 2012. (Krishnanathan et al., 2012)

- S.A. Billings. *Nonlinear System Identification: NARMAX, Methods in the Time, Frequency, and Spatio-Temporal Domains*. Wiley, 2013. (Billings, 2013)

A paper based on the material from Chapter 5 is currently under review,

- K. Krishnanathan, S. Anderson, S.A. Billings, and V. Kadiramanathan. Computational system identification of continuous-time nonlinear systems using ABC.

A conference poster in SB6.0 based on Chapter 7,

- K. Krishnanathan, S. Jaffe, S. Anderson, P. Wright, S.A. Billings, and V. Kadiramanathan. A systems and control approach to synthetic biology.

Finally, a paper is in preparation which uses material from Chapter 6 and 7,

- K. Krishnanathan, S. Jaffe, S. Anderson, P. Wright, S.A. Billings, and V. Kadiramanathan. The interpretation of gene reporter signal and key dynamic design properties.

1.5 Summary of contributions

The novel contributions coming from the thesis are:

- A data-driven nonlinear framework for identifying dynamic models to characterise genetic parts in biological systems (Chapter 4).
- A computational Bayesian identification framework for nonlinear continuous-time systems, which estimates the model parameter uncertainty by constructing a distribution that is used to capture cell population heterogeneity observed in experimental data (Chapter 5).
- Key design properties of the systems under investigation, which includes the explicit quantification of usage of cellular resources by the genetic parts (Chapter 7).
- Evidence that reporter cascades do have an influential effect on the dynamics of the systems under investigation. Also, the common invariant features in time-domain of the systems under investigation (same functional module linked to two different reporter cascades) are retrieved and adequately characterised (Chapter 6 and 7).

Chapter 2

Synthetic biology as an engineering problem

2.1 Introduction

The Polish geneticist Szybalski was the first to mention the new emerging field called synthetic biology in 1974 (Szybalski, 1974). Back then he envisioned that once the descriptive understanding of molecular biology is achieved, a whole new challenge in research will start. A research about devising new control genetic parts, that could be added to existing genomes for several novel ideas. The discovery of constructing recombinant deoxyribonucleic acid (DNA) using restriction enzymes in 1978, propelled the field of synthetic biology (Szybalski and Skalka, 1978). Recombinant DNA molecules are artificial synthesised DNA which consist of multiple genetic sequences that are ligated together. The recombinant DNA molecules are attached to plasmids, which are inserted into naturally occurring cells known as the host cells. The host cells acquire new functionalities through the inserted recombinant DNA.

Approximately 30 years after the discovery of recombinant DNA, synthetic biology has grown rapidly, as synthesised genetic parts have more complex and diverse functionalities. There are different synthetic biology research projects being conducted, related to various industrial sectors such as health-care (Khalil and Collins, 2010, Lu and Collins, 2009, Weber et al., 2008), food (Stocker et al., 2003, van der Meer and Belkin, 2010) and energy (Alper and Stephanopoulos, 2009, Atsumi et al., 2008, Keasling and Chou, 2008). In Keasling's group, the development of a genetic regulatory network pathway to produce an anti-malarial drug precursor, artemisinic acid in yeast, attracted attention and success (Dueber et al., 2009,

Ro et al., 2006, Shiba et al., 2007). A Paris-based pharmaceutical company Sanofi, licensed this technology in 2008 and is reported to have successfully produced 39 tonnes of artemisinic acid by 2013 (Peplow and others., 2013). Recent advancement have also seen the emergence of optogenetics, the use of certain wavelengths of light as an external mode of control towards gene regulation (Miliadis-Argeitis et al., 2011, Shimizu-Sato et al., 2002, Toettcher et al., 2011) since the breakthrough of engineering bacteria to see light in 2005 (Levskaya et al., 2005). This mode of control proves to be beneficial, as chemical interference with the host cell's cellular context is avoided and enables achievement of precise control.

This chapter discusses the engineering concepts shared in synthetic biology and the analogous design of genetic parts to various computer and electrical systems. As various similarities between engineering and synthetic biology are reviewed, the importance of practising classical engineering strategies: standardisation, decoupling and abstraction, in synthetic biology is highlighted. Challenges in designing genetic parts are also discussed with links to characterisation. Model-based design as a solution to these challenges is reflected, and existing models in the literature for both cell growth and gene expression are reviewed with their shortcomings.

2.2 An engineering problem with great complexity

Synthetic biology has gone through 3 phases: molecular, modular and system level, in terms of understanding and designing cellular functions (Hartwell et al., 1999, Purnick and Weiss, 2009). Presently, designed genetic parts in synthetic biology are analysed and tested at the system level, where application based synthesised biological systems are evolving from design prototypes (McDaniel and Weiss, 2005). Cells are made up of several different types of molecules, which interact with one another that enables the cells to perform various functionalities. In the twentieth century, biologists tried to define cellular functions at the molecular level, however, this turned out to be very complex (Hartwell et al., 1999). The reason being, each cellular function is not carried out by a single molecule nor does a single molecule have only one distinct functionality to carry out. Rather, a group of molecules interact with one another to perform a discrete cellular functionality. The group of molecules is termed a module. Cellular functions are easier to define and interpret by module or group of modules. Cells as a whole, exploit the interconnectivity of group of modules, to achieve higher level of cellular functions and operate as a system. Recombinant DNA can be classified as an artificially synthe-

sised module. The modular and system level understanding of cellular functions, aid researchers in the twenty-first century to design genetic parts in biological systems based on physical layers, that constitute a module or group of modules (Andrianantoandro et al., 2006).

Interactions between molecules in cells occur through the process of biochemical reactions. There are vast number of different biochemical reactions that take place inside a cell, which enables the design of biological systems with a variety of functionalities. Some of the biochemical reactions that take place in a cell are transcription, translation, protein phosphorylation, allosteric regulation and ligand/receptor binding (Andrianantoandro et al., 2006, Burack and Sturgill, 1997, Jewett et al., 2008). When two or more biochemical reactions are fused together, a genetic regulatory network is created which gives rise to a functional module. Some examples of genetic regulatory networks are transcriptional regulation networks, protein signalling pathways and metabolic networks. Transcriptional regulation networks are well studied and implemented (Figure 2.1). By modifying and tweaking certain properties of a transcriptional module, mechanisms such as positive and negative feedback, feedforward and multi-transcriptional cascade can be achieved. By adopting the positive and negative feedback mechanisms, functions such as: (i) cell bistability can be attained - the induction of an activator or a repressor switches a cell from one stable state to another (Berg, 1988, Maeda and Sano, 2006, Morgan, 1997, Stricker et al., 2008), (ii) a toggle switch can be constructed, by using one of each positive and negative feedback mechanisms (Gardner et al., 2000, Tian and Burrage, 2006) and (iii) an oscillatory dynamics can be achieved, by the use of several positive and negative feedback mechanisms (Atkinson et al., 2003, Elowitz and Leibler, 2000, Garcia-Ojalvo et al., 2004, Stricker et al., 2008). The feedforward mechanism is built with a single transcriptional module that is controlled by two transcription factors (Basu et al., 2005, Bleris et al., 2011, Mangan and Alon, 2003), whereas the multi-transcriptional cascade mechanism is built by connecting several transcriptional module, side by side, that could be used in studying delays in signal transmission (Hooshangi et al., 2005, Rosenfeld and Alon, 2003).

The hierarchical levels in studying and designing genetic parts in biological systems draws synthetic biology more towards engineering than natural science (Andrianantoandro et al., 2006, Endy, 2005, Hartwell et al., 1999). In Andrianantoandro et al. (2006), conceptual analogy between computer engineering and synthetic biology is shown, such as networks to tissues, computers to cells, modules to

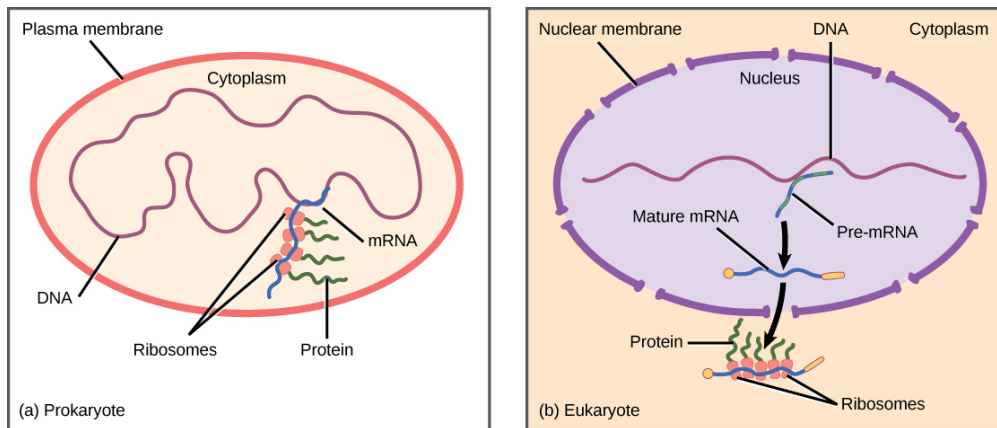


Figure 2.1: Transcription and translation steps in a) prokaryotic cells and b) eukaryotic cells (image obtained from openstax college free online book - biology). In this thesis, only bacterial cells are considered, which are prokaryotic cells. During transcription, RNA copy of a gene sequence is made, called the messenger RNA that is read by ribosomes, in order to translate it into a sequence of amino-acids during protein synthesis.

pathways *etc.*. As much comparison is made between engineering and synthetic biology, there are unique properties that distinguishes biological systems from systems in other engineering fields. These unique properties are cellular dependency, replication and evolution (Andrianantoandro et al., 2006, Heinemann and Panke, 2006). Genetic parts are cellular dependent - they typically need cellular resources to function. They obtain and share the resources from the engineered host cells, which leads to crosstalk (interference in the genetic regulatory network), thereby modifying the host cells themselves. There is recent research focusing on building and testing transcriptional-translational genetic parts in a cell-free environment (Shin, 2012, Siegal-Gaskins et al., 2013, Tuza et al.).

Replication is a naturally occurring characteristic of cells. Parent cells divide and multiply into several daughter cells during growth phase. Inserted genetic parts within the cells also divide and replicate along with the cells. Even though a cell is engineered to acquire a desired functionality, populations of cells are used to complete the desired and required tasks. Mostly, the populations of cells exhibit heterogeneity and variability in their performance causing uncertainty. However, the effect of unpredictability in the molecular level will be of minimal effect, given that a significant amount of cells in a particular population perform the desired tasks.

Evolution can be defined as the modification of the inherited characteristics in cells

over successive generations. It can be both disadvantageous and advantageous to the design of biological systems. Disadvantageous because evolution introduces variability in the characteristics of cells, leading to unpredictability and, it also optimises the wild strains of the cells to the extent it refuses to be compatible in an artificial context (Andrianantoandro et al., 2006). Directed evolution is the process of implementing related restraints to command natural selection, in order to produce desirable protein molecules. This serves as an optimisation procedure in the design of biological systems which is advantageous (Collins et al., 2005, Haseltine and Arnold, 2007, Isaacs et al., 2006, Purnick and Weiss, 2009). Amongst the synthetic biology community, the designing of genetic parts in biological systems is approached by practising classical engineering strategies: standardisation, decoupling and abstraction, with the consideration of the distinguishing properties mentioned above (Endy, 2005, Hartwell et al., 1999, Heinemann and Panke, 2006).

2.2.1 Standardisation, decoupling and abstraction

Standardisation is the definition and implementation of standards in genetic parts. Standardisation ensures that genetic parts are designed in a defined manner, that permits the fabrication of higher order genetic parts to be translatable and uncomplicated. This motivates non-experts such as engineers, to get involved in the design process. In Müller and Arndt (2012), it states that:

"standardization of the physical composition and the description of each part is required as well as a controlled vocabulary to aid design and ensure interoperability."

There are two well renowned establishments which provide standard genetic parts: (i) registry of standard biological parts (RSBP) (http://parts.igem.org/Main_Page?title=Main_Page) and (ii) biofab (<http://www.biofab.org/>). RSBP was founded at the Massachusetts Institute of Technology (MIT), its registry holds around 6000 standard genetic parts (Shetty et al., 2011). The standard genetic parts are annually contributed into the registry by the international genetic engineering machine (IGEM) teams and laboratories. IGEM is an annual competition which attracts laboratories and teams from different parts of the world to educate and compete, in order to advance the field of synthetic biology, and develop an open community and collaboration. The standard of a genetic part in RSBP is defined by the use of a prefix and suffix, in the beginning and end of a genetic part. The prefix and suffix segments contains restriction sites, that could be cut and connected using specific restriction enzymes and DNA ligase, permitting the assembly of a new genetic part. The assembly also ensures that the new genetic part maintains the same prefix and suffix segments.

Registry of Standard Biological Parts

- Browse [chassis](#)
- Browse [user-supplied catalog pages](#) - these pages have not undergone curation. Please feel free to add new catalog pages to this section.

Browse parts by type













Catalog	List
	 Promoters (?) : A promoter is a DNA sequence that tends to recruit RNA polymerase to the start of the downstream DNA sequence.
	 Ribosome Binding Site/about (?) : A ribosome binding site (RBS) is a sequence of RNA that ribosomes can bind and initiate translation.
	 Protein domains (?) : Protein domains are portions of proteins cloned from a protein coding sequence. Some protein domains might change the function of the protein for cleavage, or enable it to be readily purified.
	 Protein coding sequences (?) : Protein coding sequences encode the amino acid sequence of a protein from start codon to stop codon. Coding sequences for genes and cDNAs are also included here.
	 Translational units (?) : Translational units are composed of a ribosome binding site, a start codon, and a stop codon. They begin at the site of translational initiation, the RBS, and end at the stop codon.
	 Terminators (?) : A terminator is an RNA sequence that usually occurs at the end of a gene and causes transcription to stop.

Figure 2.2: A snapshot of the catalogue in the registry of standard biological parts

The plasmid backbone which propagates the new genetic part, also defines the standard of the new genetic part it maintains. RSBP holds a diverse catalogue of genetic parts (Figure 2.2), such as promoters, terminators, ribosome binding sites, translational units *etc.*

Synthetic biology requires expertise from different fields, to work together to produce a functioning whole. Decoupling plays a crucial role in this by separating a complicated problem into several simpler problems, that can be directed to individual experts, such that the combination of the resulting work could produce the required solution (Bashor et al., 2010, Endy, 2005, Tucker and Zilinskas, 2006). A good example is the DNA synthesis, where certain experts focus on designing useful pieces of DNA, while others focus on using the DNA pieces to build DNA.

In the engineering of a biological system, when decoupling is achieved, the respective simpler tasks which make up the main goal can be organised across levels of complexity using abstraction hierarchies. This helps in managing engineering complexity. Working on each level should be independent and transparent. Some

current computer design tools support hierarchical design of biological systems, aiding redesign and simplicity (Chandran et al., 2009).

2.2.2 Characterisation and model-based design

As engineering of biological systems are discussed, the daunting knowledge gap when it comes to how cells operate is a barrier. The difficulties multiply as the genetic regulatory networks get larger, limiting the ability to design more complex biological systems. In 2009, the review Purnick and Weiss (2009), shows that although the number of synthesised genetic parts has risen over the past few years, the complexity of their genetic regulatory networks has begun to flatten out. Challenges are encountered in every step of the process, from the characterisation of genetic parts to the fabrication of the biological systems. In Kwok (2010), five challenges are reported: (i) many genetic parts are undefined, (ii) the genetic regulatory network is unpredictable, (iii) the complexity is unwieldy, (iv) many genetic parts are incompatible and (v) variability in the performance of the genetic parts crashes the biological systems. Due to these challenges, most genetic parts are built on several trial and error methods. This practice does not guarantee biological systems to be designed optimally, in a short time scale or into a working prototype.

As discussed in Chapter 1, it is important to implement a model-based design to assist the design stages of genetic parts in biological systems which would provide solutions to some challenges discussed in the above paragraph. For model-based design to be effective, deriving useful models to characterise genetic parts is crucial. However, what is fully expected from the characterization of a genetic part to aid design is still a developing process. A good example was set in Canton et al. (2008), where the emergence and importance of a datasheet, which serves as a catalogue with adequate information about a genetic part is shown. Also, as practiced in other engineering disciplines, the development of a model is emphasised and included in the datasheet. The work shown in Canton et al. (2008), serves as a benchmark for the characterisation of genetic parts and do need progressive development, for example the derived model-type (Arkin, 2008): (i) what type of model is required, (ii) what information should the model provide, (iii) should the model be static or dynamic, (iv) should the model be deterministic or stochastic, (v) how parameter constrained is the model and (vi) should it be a single-cell or population-level model? These are the very questions that the discipline of system identification tries to answer.

2.2.3 Variation in gene expression

A derived model, should be able to account for the variability observed due to population-level heterogeneity and gene noise. Biological systems dynamics are highly stochastic due to the variability, which makes modelling harder (Cheong et al., 2010, Elowitz et al., 2002, Raser and O'Shea, 2005, Swain et al., 2002, Wilkinson, 2009). Population-level heterogeneity occurs mostly due to cell death, crosstalk, mutations and, changing intracellular and extracellular conditions. Gene noise can be categorised into intrinsic and extrinsic noise at single-cell level. Intrinsic noise can be defined as an intracellular disturbance in a cell due to variations in cell resources such as RNA polymerase, transcription binding factor, messenger RNA, *etc.*. Intrinsic noise is understood to be a transient variation at the beginning of the cell cycle. Extrinsic noise arises due to cell to cell differences such as cell cycle stage, spatial chemical concentration, inheritance *etc.*. Extrinsic noise is known to last throughout the cell cycle, periodically and in small magnitude.

It is only possible to further understand gene noise, if single-cell measurement is feasible and easily attainable, which is not completely achievable presently. There are some advances in attenuating noise at single-cell level. Feedbacks and gene regulations are built as modules into cells to reduce the level of noise expression, however, the limit of suppression that is possible is very low due to the sacrifice of cells to preserve cellular resources (Lestas et al., 2008, 2010, Sun and Becskei, 2010). Also when there is sufficient resources present, the feedback signals are very noisy, which demands more cellular resources for attenuation.

2.3 Biochemical modelling

2.3.1 Cell growth models

Modelling of the microbial cell growth is important. The cell growth model should be able to capture and predict all four phases of the microbial cell growth (Figure 2.3): (i) lag, (ii) exponential, (iii) stationary and (iv) death. One of the earliest developed model for cell growth is the Monod model (Monod, 1949),

$$\mu(t) = \mu_{max} \left(\frac{s(t)}{k_{sat} + s(t)} \right), \quad (2.1)$$

where $\mu(t)$ is the changing growth rate of the microbial culture, μ_{max} is the maximum growth rate of the microbial culture, $s(t)$ is the concentration of the limiting

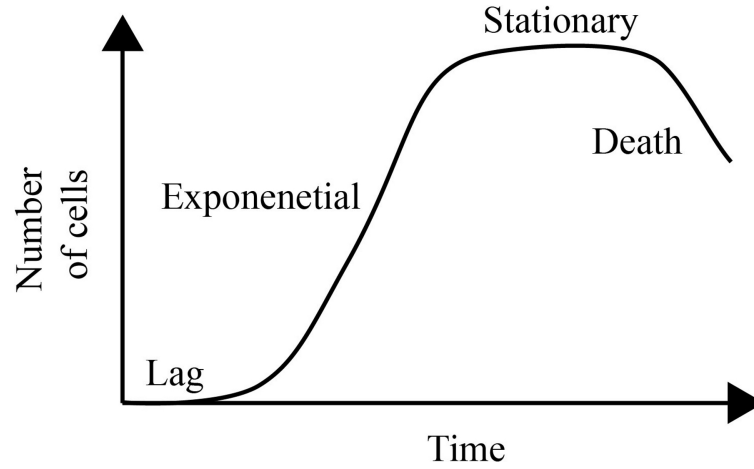


Figure 2.3: The typical growth of number of cells in a microbial culture

substrate for growth at time t and k_{sat} is the saturation constant. The Monod model only fits the exponential phase of the microbial cell growth, thereby it is unable to predict the remaining 3 phases of the microbial cell growth. There have been efforts to expand the Monod model to capture the growth of microbial cultures that consist of competitive and noncompetitive substrates and products. Another widely used model is the Logistic model that relates the changes of growth rate with changes of cell concentration. The structure of the Logistic model is defined as (Ai et al., 2003, Fujikawa et al., 2004)

$$\mu(t) = \mu_{max}x(t)\left(1 - \frac{x(t)}{x_{max}}\right), \quad (2.2)$$

where $x(t)$ is the cell concentration and x_{max} is the maximum cell concentration at the stationary phase. However, the Logistic model is unable to predict the lag phase of the microbial cell growth.

The work reported in Lin et al. (2000), develops a cell growth model (called the Lin's model in this thesis) based on the time dependent changes of growth rate $\mu(t)$, which is able to predict the lag, exponential and stationary phase of the microbial culture,

$$\mu(t) = \mu_{max}\left(\frac{1}{1 + e^{-k_{in}(t-t_{in})}}\right)\left(\frac{1}{1 + e^{k_{de}(t-t_{de})}}\right), \quad (2.3)$$

$$\dot{x}(t) = \mu(t)x(t), \quad (2.4)$$

where the maximum increasing rate of $\mu(t)$ is k_{in} , the maximum decreasing rate

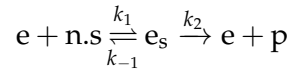
of $\mu(t)$ is k_{de} , the time point when the increasing rate of $\mu(t)$ equals k_{in} is t_{in} and the time point when the decreasing rate of $\mu(t)$ equals k_{de} is t_{de} . The parameters that govern the Lin's model can be obtained graphically from experimental data.

2.3.2 Gene expression models

There are several different types of models in the literature used in fitting gene expression data. Gene expression models can be categorised as either static or dynamic models. They are derived from reaction schemes and genetic regulatory networks which describes the underlying properties of the genetic parts under consideration. Reaction schemes and genetic regulatory networks are well studied knowledge from biochemistry and microbiology literature (Cornish-Bowden, 2013).

Static gene expression models

Static models are widely used in synthetic biology, in the form of the Michaelis-Menten and Hill equations (English et al., 2005, Gertz et al., 2008, Houston and Kenworthy, 2000, Kim et al., 2006). They are used as gray-box models to characterise input-output behaviour of modular genetic parts. Consider the simple reaction scheme used in deriving both the Michaelis-Menten and Hill equations



where e denotes enzyme, s substrate, n Hill coefficient, e_s enzyme-substrate complex, p product and where the reaction scheme has an associated total enzyme concentration e_0 . The parameters k_1 , k_{-1} and k_2 defines the rate of reactions. The ordinary differential equations (ODEs) obeying the law of mass action for the above reaction scheme can be written as (Palsson, 1987)

$$\dot{s}(t) = n(-k_1 e(t) s^n(t) + k_{-1} e_s(t)), \quad (2.5)$$

$$\dot{e}(t) = -k_1 e(t) s^n(t) + (k_1 + k_2) e_s(t), \quad (2.6)$$

$$\dot{e}_s(t) = k_1 e(t) s^n(t) - (k_1 + k_2) e_s(t), \quad (2.7)$$

$$\dot{p}(t) = k_2 e_s(t). \quad (2.8)$$

By applying assumptions which are discussed in Chapter 4 to the above ODEs,

the Hill equation can be defined as (Cornish-Bowden, 2013)

$$\dot{p}(t) = v_{max} \left(\frac{s_{ss}^n}{s_{ss}^n + k_p^n} \right), \quad (2.9)$$

where $k_p^n = \frac{k_{-1} + k_2}{k_1}$, $v_{max} = k_2 e_0(t)$ and subscript *ss* refers to steady state. The Michaelis-Menten equation can be obtained from the Hill equation by equating $n = 1$. The regressed output of the Hill equation is the vector of the rate of product at a single time point for different substrate concentration, hence called a static model. The Hill equation is modified (as the reaction schemes differ) to characterise different bimodality functions such as AND, NAND and NOT gates (Tamsir et al., 2010, Wang et al., 2011).

Dynamic gene expression models

Here, dynamic gene expression models in synthetic biology shall be reviewed by classifying them as: (i) fine-grained dynamic (FGD) models, modelling every single biochemical reactions that take place in a genetic part and (ii) coarse-grained dynamic (CGD) models, where only the input-output behaviour of a modular genetic part is characterised.

FGD modelling was once a domain mostly practised by experts, but has been an area of growth over the last decade. There has been an incremental rise in the number of software tools and open source information, facilitating model exchange and improvement. In FGD modelling, the following is needed: (i) a detailed understanding of the genetic part and its biochemical processes and (ii) encoding the biological knowledge into mathematical forms (mostly as reaction schemes or ODEs). Thereby, an initial FGD model is normally created which is reasonably faithful to the genetic regulatory network structure describing the genetic part. Mostly when implementing the FGD model of a genetic regulatory network in computer software, it is done in many logical layers (Figure 2.4B), where the first layer comprises of interacting molecules such as messenger RNA, RNA polymerase, proteins *etc.*. The final layer comprises of the mathematical expressions (in form of reaction schemes or ODEs) for describing the biochemical interactions, and the fluxes in and out of the biological systems.

The extensible markup language (XML) formats for FGD models of genetic regulatory networks have enabled the use of different computer programs and tools, and the two commonly used extensions are systems biology markup language (SBML) and cellular markup language (CellML) (Hucka et al., 2003, Lloyd et al.,

2004). These extensions allow the use of own internal representation of a mathematical model (not restricted to ODEs). The scripting language in the extensions will entitle the following elements such as species, parameters, reactions, rules, events *etc.*, to describe the genetic regulatory network. Some of the biochemical and gene regulatory interactions with kinetic parameters can be obtained from scientific databases (Caspi et al., 2008, Kanehisa et al., 2008, Vastrik et al., 2007). Unidentified parameters are mostly estimated using a range of global and local optimisation algorithms offered by COPASI (Hoops et al., 2006) or inspected qualitatively by bifurcation analysis (Endler et al., 2009).

CGD modelling in synthetic biology is also widely practised, which involves the characterisation of input-output behaviour for a modular genetic part. Apart from the static gene expression models such as Michaelis-Menten and Hill equations which are used as gray-box models, dynamic gray-box models under CGD modelling are also explored as deterministic and stochastic functions in the form of ODEs and stochastic differential equations (SDEs) (MacDonald et al., 2011). CGD models are mostly predefined structural models, which are usually incorporated with a sigmoidal function to capture the cooperative binding of ligands, however, their structures largely depend on the genetic part under consideration. Some good examples of CGD models are the well known repressilator model (Elowitz and Leibler, 2000) and models reported in (Covert et al., 2008, Karlebach and Shamir, 2008, Tian and Burrage, 2006, Wilkinson, 2009) (Figure 2.4A). If a statistical approach is taken, the parameter estimation of the CGD models are done by maximum likelihood method, whereas, weighted sum of squares or other optimisation algorithms are used for non-statistical approach (De Jong, 2002, MacDonald et al., 2011) (see Chapter 3). Model selection has been recently explored for CGD modelling, when competing CGD models are involved. The competing CGD models are nested and compared using a computational Bayesian method called the approximate Bayesian computation (ABC) (Barnes et al., 2011, Kirk et al., 2013) (see Chapter 5).

Shortcomings in gene expression models

The modelling process for gene expression itself allows one to scrutinise not only the available experimental data, but also the known or assumed models. Therefore, to conclude this subsection, the shortcomings of existing gene expression models which have evidently resulted to ineffective model-based design of genetic parts are outlined:

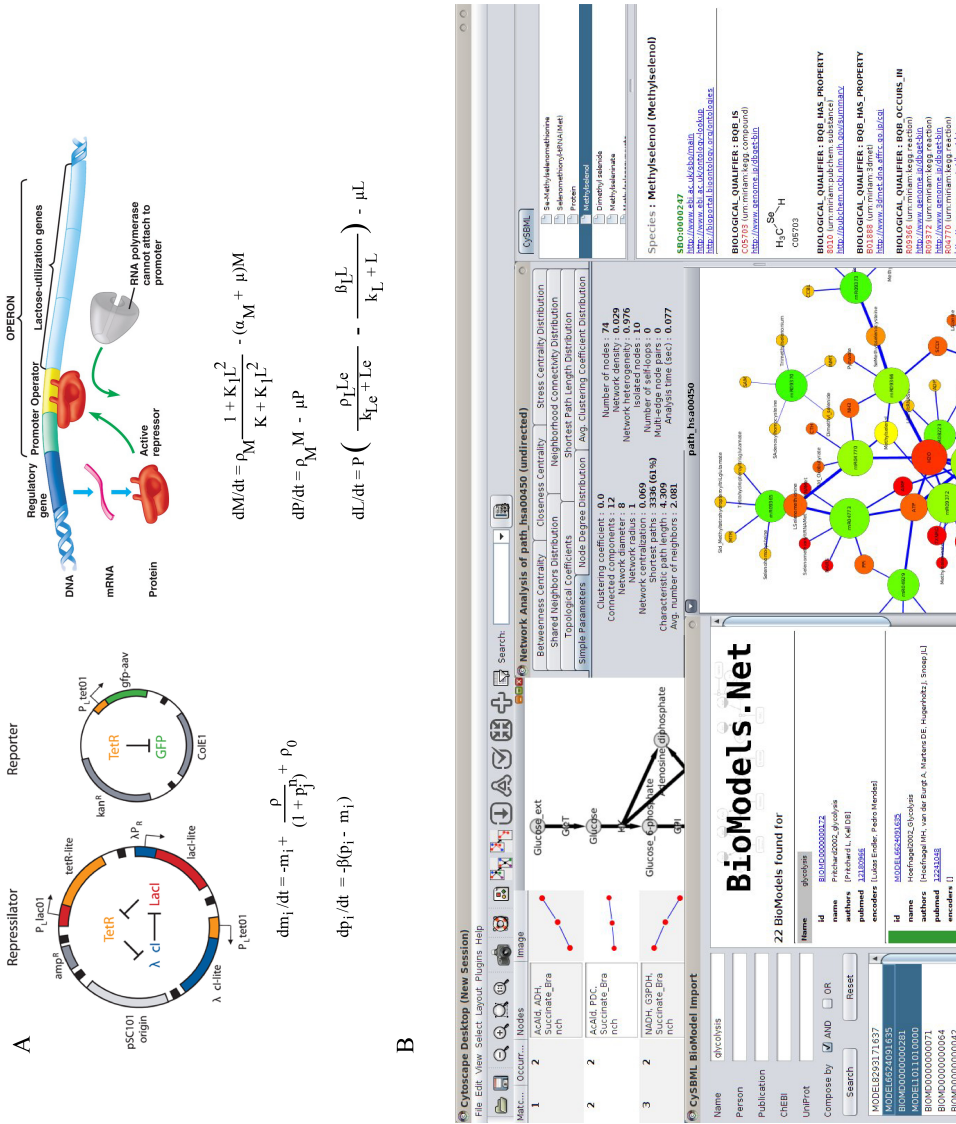


Figure 2.4: Examples of CGD models. A. Models of the repressor (left) (Elowitz and Leibler, 2000) and Lac operon (right) (Lestas et al., 2008), and B. A snapshot of the CysBML, which provides a platform to import SBML files ([http:// apps.cytoscape.org/ apps/ cysbml](http://apps.cytoscape.org/apps/cysbml)).

- The static models are shown to have good experimental fit and predictions whereas most dynamic models are not. However the static models are limited in terms of design implementation.
- Complexity is a major issue in synthetic biology. The dynamic models grow in number as the genetic regulatory network becomes larger leading to an explosion in model complexity. High level of detailed resolution in modelling is both computationally intractable and too quantitatively inaccurate to answer the questions that one is interested in, during the design of genetic parts.
- Most of the static and dynamic models have predefined model structure, that is not adjusted to the complexity of the genetic part under consideration. Due to this, most parameters of the models are constrained and leads to unrepeatability towards another set of experimental data.

2.4 Summary

This chapter gave an overview of the similarities between synthetic biology and engineering. It also discussed the inherent complexity involved in engineering genetic parts in biological systems, and the need of implementing model-based design to assist the design stages of biofabrication. Cell growth and gene expression models in the literature were reviewed, with the latter having shortcomings due to model complexity, poor experimental data prediction and computational intractability for design.

Chapter 3

System identification and its literature

3.1 Introduction

As discussed in the previous two chapters, model-based design of genetic parts in biological systems, has not been effective due to the shortcomings of dynamic gene expression models in the synthetic biology literature. In this thesis, we address this problem by introducing nonlinear black-box models in system identification. Nonlinear black-box models are able to encapsulate and capture all internal functions of a genetic part, thereby characterising it using only the input and output data. In later chapters, this is shown to be sufficient for knowing how to use this abstracted genetic part in a larger design. The nonlinear black-box models are suggested to overcome the shortcomings of existing dynamic gene expression models that are overly complex, unwieldy, and of unknown structure (Kwok, 2010).

The acquisition of input-output data through the design of experiments facilitates the identification of black-box models, which requires the input signal to persistently excite the system to evoke its full range of dynamics, that is observed in the output data (Ljung, 1999). There is an abundance of black-box model structures under different model class: linear, nonlinear, discrete-time, continuous-time, parametric, non-parametric, time-variant, time-invariant, *etc.* (Pearson, 2003). In this chapter, linear and nonlinear in model structure, discrete-time and continuous-time, time-invariant parametric models that are linear (nonlinear autoregressive model with exogenous input) and nonlinear (nonlinear output error model) in-the-parameters are reviewed.

Since parametric models are used, parameter estimation is a key step, least squares is used in estimating linear parameters (Ljung, 1999) and gradient-based nonlinear optimisation is used in estimating nonlinear parameters (Nelles, 2001). The experimental data modelled in this thesis is nonlinear, therefore the challenging task of model structure detection is discussed. The model structure detection to identify a nonlinear black-box model with parsimonious description can be approached either by forward or backward selection of model terms from a superset of possible candidate model terms. The model structure detection methods reviewed here are shown to be driven by: (i) the error reduction ratio, either computed from one-step-ahead (Chen et al., 1989) or simulated (Piroddi and Spinelli, 2003) prediction or (ii) using estimates of parameter statistics to selectively include or exclude model terms (Kukreja et al., 2004). In system identification, the model identified is expected to have a good generalisation performance, which is accessed by cross-validating the model with the test data. The parametric nonlinear black-box models can be transformed directly from time domain models to frequency domain models using the concept of generalised frequency response function (Billings and Tsang, 1989a), which were introduced to describe the spectral properties of nonlinear dynamical systems.

This chapter deals with the concept of modelling dynamical systems from observed input and output data. Parametric models with a superset of possible model terms are used, so estimation of the parameter values and detection of correct model terms is of fundamental importance. The nonlinear black-box models, modelling techniques and the frequency domain analysis based on generalised frequency response function presented in this chapter represent a comprehensive package of tools that are used in the subsequent chapters to model and analyse the genetic parts of biological systems.

3.2 Model structures

The model structures discussed in this section are formulated in discrete-time (DT). It should be noted that, the models implemented in this thesis are in CT (Chapter 4,5 and 7), however, as system identification is much more established in DT, there is a need to review both DT and CT model structures. In section 3.7, some equivalent CT model structures are later discussed briefly. There is a wide variety of model structures that are available which are not discussed in this thesis such as Wiener series, Volterra series *etc.*, as it exceeds the scope of this overview and the reader is referred to reviews and books available (Billings, 2013,

1980, Haber and Unbehauen, 1990, Nelles, 2001).

3.2.1 Linear black-box models

The representation of the generic black-box model structure for linear systems, which is a widely accepted standard in system identification is given in Ljung (1999),

$$y_k = \frac{B(q)}{F(q)A(q)}u_k + \frac{C(q)}{D(q)A(q)}e_k, \quad (3.1)$$

where $y_k \in \mathbb{R}$, $u_k \in \mathbb{R}$ and e_k are the discretised output, input and noise signal at time step k respectively. The noise e_k is assumed to be independent, zero-mean and white. The operator q denotes the forward shift, *i.e.*, $q^{-1}y_k = y_{k-1}$. The input and noise transfer functions can be further simplified by representing them with the filters

$$G(q) = \frac{B(q)}{F(q)A(q)} \text{ and} \quad (3.2)$$

$$H(q) = \frac{C(q)}{D(q)A(q)}. \quad (3.3)$$

The observed output y_k normally contains additive noise e_k , which is mostly inherent from the process or due to sampling errors *etc.*. All extraneous behaviour is assumed to be included in the disturbance term which consists of a rational transfer function $H(q)$ driven by e_k . Two commonly used noise models are

$$y_k = \frac{1}{D(q)}e_k \text{ and} \quad (3.4)$$

$$y_k = C(q)e_k, \quad (3.5)$$

where eqn(3.4) is the autoregressive (AR) model and eqn(3.5) is the moving average (MA) model.

Some common linear black-box models are shown in Table (3.1) (Nelles, 2001). The autoregressive model with exogenous input (ARX) is the simplest linear black-box model whose prediction error is linear-in-the parameters. The autoregressive moving average model with exogenous input (ARMAX) allows for adequate flexibility in describing the properties of the noise e_k as a MA noise model. However, the ARMAX model is nonlinear-in-the parameters, therefore requires a more

Table 3.1: Some common linear black-box models

Model structures	Model equations
ARX	$y_k = \frac{B(q)}{A(q)}u_k + \frac{1}{A(q)}e_k$
ARMAX	$y_k = \frac{B(q)}{A(q)}u_k + \frac{C(q)}{A(q)}e_k$
OE	$y_k = \frac{B(q)}{F(q)}u_k + e_k$

computation demanding parameter estimation than expected of the ARX model (nonlinear parameter estimation). A more straightforward linear black-box model is the output error (OE) model, where the noise e_k is assumed to directly disturb the system process additively at the output y_k . This is a mostly assumed to be a more realistic model to describe real physical systems.

3.2.2 Nonlinear black-box models

The transition from linear to nonlinear black-box model structures is discussed below. For simplicity, the nonlinear black-box model structures are formulated for single-input single-output (SISO) systems. The ARX model can be extended to a nonlinear model structure as a nonlinear autoregressive model with exogenous input (NARX) model (Leontaritis and Billings, 1985a,b), which can be represented by the difference equation below

$$y_k = f(y_{k-1}, \dots, y_{k-n_y}, u_{k-1}, \dots, u_{k-n_u}) + e_k, \quad (3.6)$$

where n_y and n_u denote the maximum number of lags in the discrete output and input signal, and $f(\cdot)$ is a mapping function that describes the dynamics of the nonlinear process. Normally, the functional structure of $f(\cdot)$ is usually not known, however, various expansions have been studied that can arbitrary well approximate $f(\cdot)$. The output y_k and input u_k signals are time-series data obtained by sampling continuous-time data $y(t)$ and $u(t)$ at a sampling time T in the interval $t_k = kT$ for $k = 0, \dots, N_y - 1$ where N_y is the number of data samples.

The nonlinear generalisation of the ARMAX model is the nonlinear autoregressive moving average model with exogenous input (NARMAX) (Leontaritis and Billings, 1985a,b) that can be represented as

$$y_k = f(y_{k-1}, \dots, y_{k-n_y}, u_{k-1}, \dots, u_{k-n_u}, e_{k-1}, \dots, e_{k-n_e}) + e_k, \quad (3.7)$$

where n_e denotes the maximum number of lag in the noise signal. Similar to the difference in the linear case (the comparison between the ARX and ARMAX model), the NARX model does not have a noise model whereas the NARMAX model does.

The last nonlinear black-box model structure to be reviewed is the nonlinear output error (NOE) model (Nelles, 2001),

$$z_k = f(z_{k-1}, \dots, z_{k-n_z}, u_{k-1}, \dots, u_{k-n_u}), \quad (3.8)$$

$$y_k = z_k + e_k, \quad (3.9)$$

where $z_k \in \mathbb{R}$ is the undisturbed discretised output signal at time step k and n_z denotes the maximum number of lag in the undisturbed output signal.

The nonlinear function $f(\cdot)$ can be decomposed and represented by a linear sum of basis functions $\phi_j(\cdot)$, which can have varying forms including wavelet, polynomial or radial functions,

$$f(\mathbf{S}_k) = \sum_{j=1}^{N_\theta} \theta_j \phi_j(\mathbf{S}_k), \quad (3.10)$$

where N_θ is the number of model terms, θ_j is the parameter associated with basis function $\phi_j(\cdot)$ and \mathbf{S}_k is a vector of lagged variables which is dependent on the nonlinear black-box model structure. In the order of NARX, NARMAX and NOE models, their corresponding \mathbf{S}_k vector can be represented as

$$\text{NARX: } \mathbf{S}_k = (y_{k-1}, \dots, y_{k-n_y}, u_{k-1}, \dots, u_{k-n_u}), \quad (3.11)$$

$$\text{NARMAX: } \mathbf{S}_k = (y_{k-1}, \dots, y_{k-n_y}, u_{k-1}, \dots, u_{k-n_u}, e_{k-1}, \dots, e_{k-n_e}) \text{ and} \quad (3.12)$$

$$\text{NOE: } \mathbf{S}_k = (z_{k-1}, \dots, z_{k-n_z}, u_{k-1}, \dots, u_{k-n_u}) \quad (3.13)$$

respectively.

Advantages of nonlinear black-box models: NARX, NARMAX and NOE

Many of commonly applied models implemented to characterise genetic parts in synthetic biology have limitations, either on the model representation and complexity, model prediction performance or system design. In contrast, the nonlinear black-box models (discussed above) provide a general solution that can address all challenges in modelling of genetic parts in biological systems. The advantages can be summarised by the following points:

- The reviewed nonlinear black-box models can be used to represent a wide range of nonlinear dynamical systems.
- The reviewed nonlinear black-box model's representations are parsimonious, which is aided by the automated data-driven model structure detection (MSD) which allow the identification of very sparse system descriptions involving a small number of parameters (see section 3.4). This is in comparison to alternatives such as the Volterra series and biochemically derived ordinary/stochastic differential equations.
- The reviewed nonlinear black-box models can be directly transformed into the frequency domain as generalised frequency response functions (GFRFs) (see section 3.8). Thereby, allowing for an integrated methodology for identification, analysis and design.

3.2.3 Cascade models

Cascade models are widely used to describe nonlinear systems, more often in biology (Bai, 2002, Gollisch and Meister, 2008, Westwick and Kearney, 2001), due to their relative simplicity in physical interpretations and the ability to preserve the system's structure (Billings, 1980). Each cascade block consists of elementary processing units which perform either dynamic linear operations (L) or time independent, static nonlinear operations (N); thereby separating the nonlinearity from the dynamics of the system's process (Nelles, 2001). The two simplest and perhaps most commonly used cascade models are the Hammerstein and Wiener models (Figure 3.1). A Hammerstein model consists of a single N-L cascade blocks, which can be represented by the equations

$$v_k = f_s(u_k), \quad (3.14)$$

$$y_k = \sum_{j=-\infty}^{\infty} v_{k-j} h_j, \quad (3.15)$$

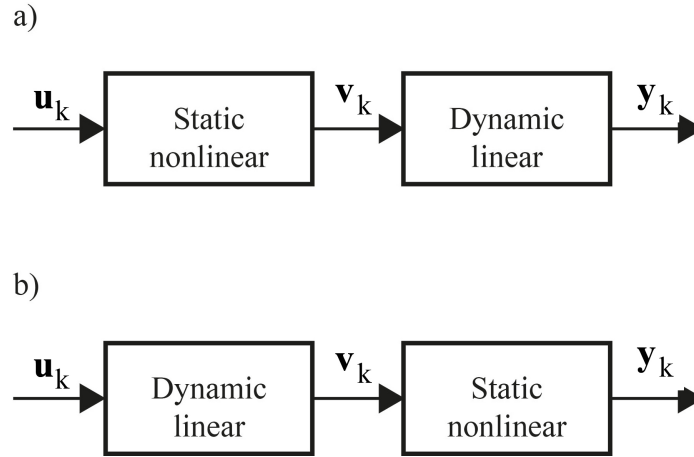


Figure 3.1: Cascade models a) Hammerstein model and b) Wiener model.

where $f_s(\cdot)$ is the static nonlinearity, which can be represented using either polynomial, radial or wavelet functions (Pearson, 1999), and h_j denotes the filter's impulse response function. Whereas the Wiener model corresponds to a single L-N cascade blocks,

$$v_k = \sum_{j=-\infty}^{\infty} u_{k-j} h_j, \quad (3.16)$$

$$y_k = f_s(v_k). \quad (3.17)$$

The estimation of the parameters for each cascade block is done independently in an iterative procedure. The range of methods available exceeds the scope of this overview and the reader is referred to reviews available (Giri and Bai, 2010, Hunter and Korenberg, 1986). The separable least squares (SLS) method which is reviewed in section 3.3.2, can also be used in estimating parameters of cascade models, which is shown to have a better performance than the existing iterative procedures (Westwick and Kearney, 2001).

Simple cascade models can be extended to broader types of cascade models such as the Hammerstein-Wiener (N-L-N) model (Bai, 2002) and L-N-L cascade model (Korenberg and Hunter, 1986). A challenging problem in cascade models is that intermediate signals such as v_k are in most cases not available. The estimation problem in broader cascade models are also nonlinear-in-the parameters which could be solved numerically by nonlinear optimisation, provided the exact model structure of each cascade blocks are known (Niven et al., 2003).

To conclude this section, after an appropriate model structure is chosen to represent a system's dynamics accordingly, the following step would be the model estimation. The model estimation in system identification, is a combined problem of MSD and parameter estimation. MSD involves identifying a model with parsimonious representation such that it will fit the training data set well but also general enough for other experimental datasets. Parameter estimation on the other hand is the problem of estimating unbiased parameters for the selected model terms.

3.3 Parameter estimation

In this section, a general introduction to parameter estimation which entails different optimisation methods that allow one to determine the model's optimal parameters are discussed. The full literature of parameter estimation available exceeds the scope of this overview and the reader is referred to books available (Bar-Shalom et al., 2004, Bishop and Nasrabadi, 2006, Ljung, 1999, Nelles, 2001, Söderström and Stoica, 1988). The optimisation methods discussed here, will be those that are classed under the supervised learning, supervised learning are implemented based on the knowledge about the input and output data of the system.

The models dealt with in this thesis are both linear and nonlinear -in-the-parameters, where the parameter themselves are time-invariant. Given the input vector \mathbf{u} and measurement output vector \mathbf{y} , parameter estimate $\hat{\boldsymbol{\theta}}$ needs to be obtained that best represent the measurement data \mathbf{y} ,

$$\mathbf{u} = \left(u_0, u_1, \dots, u_{N_y-1} \right)^\top = \left(u(t_0), u(t_1), \dots, u(t_{N_y-1}) \right)^\top, \quad (3.18)$$

$$\mathbf{y} = \left(y_0, y_1, \dots, y_{N_y-1} \right)^\top = \left(y(t_0), y(t_1), \dots, y(t_{N_y-1}) \right)^\top, \quad (3.19)$$

$$\boldsymbol{\theta} = \left(\theta_1, \dots, \theta_{N_\theta} \right)^\top. \quad (3.20)$$

The optimisation methods will be reviewed by breaking it further down to statistical and non-statistical approaches. It is useful and instructive to briefly describe basic aspects of both approaches and relate them to system identification,

1. Statistical approach: if the assumption that there is an unknown true value of $\boldsymbol{\theta}$ is taken, maximum likelihood (ML) maximises how likely a parameter is, given the observation \mathbf{y} that have been made. The maximum a posteriori

(MAP) takes a conceptually different treatment towards parameter estimation which shall be discussed below.

2. Non-statistical approach: this will be discussed under 2 sub-categories, linear and nonlinear methods. The nonlinear methods discussed here will be focused on basic nonlinear local optimisation which is related to the gradient-based techniques. A brief review of SLS is also discussed, which utilises both linear and nonlinear methods.

If a model with linear-in-the-parameters is assumed, a linear regression model can be formulated to represent it,

$$y_k = \boldsymbol{\phi}_k^\top \boldsymbol{\theta} + e_k, \quad (3.21)$$

where $\boldsymbol{\phi}_k$ is the regression vector containing regressors. For a given estimate $\hat{\boldsymbol{\theta}}$, the associated output prediction will be

$$\hat{y}_k = \boldsymbol{\phi}_k^\top \hat{\boldsymbol{\theta}}, \quad (3.22)$$

therefore the prediction error can be described as

$$\hat{e}_k = y_k - \hat{y}_k. \quad (3.23)$$

3.3.1 Statistical approach

Maximum likelihood

The likelihood function is defined by $p(\mathbf{y}|\boldsymbol{\theta})$ and the maximum likelihood method seeks to maximise this likelihood function, such that

$$\hat{\boldsymbol{\theta}}_{ML} = \arg \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}). \quad (3.24)$$

Considering eqn(3.21) and eqn(3.22) again, with the added assumption that e_k is normally distributed with variance λ^2 , then

$$p(\mathbf{y}|\Phi; \boldsymbol{\theta}) = \prod_{k=0}^{N_y-1} p(y_k|\boldsymbol{\phi}_k; \boldsymbol{\theta}), \quad (3.25)$$

where

$$p(y_k|\boldsymbol{\phi}_k; \boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\phi}_k^\top \boldsymbol{\theta}, \lambda^2). \quad (3.26)$$

Logarithm functions are monotonically increasing functions, therefore minimising a function is equivalent to minimising its logarithm. In order to simplify eqn(3.26), logarithms can be taken on both sides and differentiated with respect to θ . The following equation below provides the solution

$$\hat{\theta}_{ML} = \left(\sum_{k=0}^{N_y-1} \phi_k^\top \phi_k \right)^{-1} \left(\sum_{k=0}^{N_y-1} \phi_k^\top y_k \right). \quad (3.27)$$

Therefore when assuming Gaussian additive noise, maximising the likelihood function will be equivalent to minimising the sum of errors squared (shown below).

Maximum a posteriori

In a Bayesian framework, the parameters θ itself are thought of as a random variable. Based on the observation \mathbf{y} , which is another random variable that is correlated with the parameters, we may infer information about their value. Assuming a prior distribution of the parameters $\pi(\theta)$ exists and using Bayes' rule, the posterior distribution of the parameters $p(\theta|\mathbf{y})$ can be obtained as

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)\pi(\theta)}{p(\mathbf{y})}. \quad (3.28)$$

where $p(\mathbf{y})$ is the marginal likelihood, which in practice is mostly unknown and impossible to calculate.

The Bayesian framework estimates a distribution over the parameters of a model. With the distribution over the parameters it is possible to integrate over the possibilities to get an average prediction for future outcomes of the system and full sense of the uncertainties that pertain to it. Also from the posterior distribution, different estimates of θ can be determined, for example, the parameter set corresponding to the maximum value the distribution attains, which is also referred as the MAP estimate,

$$\hat{\theta}_{MAP} = \arg \max_{\theta} \left(p(\mathbf{y}|\theta)\pi(\theta) \right). \quad (3.29)$$

3.3.2 Non-statistical approach

Linear methods

Considering eqn(3.23), the least squares (LS) estimate $\hat{\theta}_{LS}$ is the value of θ which minimises the sum of errors squared,

$$\hat{\theta}_{LS} = \arg \min_{\theta} \sum_{k=0}^{N_y-1} \hat{e}_k \hat{e}_k^{\top}. \quad (3.30)$$

Using eqn(3.21) and eqn(3.22), the cost function to be minimised with respect to θ can be written as

$$J(\theta) = \sum_{k=0}^{N_y-1} \left(y_k - \phi_k^{\top} \hat{\theta} \right) \left(y_k - \phi_k^{\top} \hat{\theta} \right)^{\top}, \quad (3.31)$$

which can be expanded, rearranged, differentiated and equated to zero to show that it is minimised when

$$\hat{\theta}_{LS} = \left(\sum_{k=0}^{N_y-1} \phi_k^{\top} \phi_k \right)^{-1} \left(\sum_{k=0}^{N_y-1} \phi_k^{\top} y_k \right), \quad (3.32)$$

which can also be written in the matrix form as

$$\hat{\theta}_{LS} = \left(\Phi^{\top} \Phi \right)^{-1} \Phi^{\top} \mathbf{y}. \quad (3.33)$$

If a more realistic situation is considered, where a system is corrupted by additive noise that is correlated (such as the OE and NOE models), then the LS will give biased estimates of the parameters. The generalised least squares (GLS) overcomes this problem to provide an unbiased estimate of the parameters by

$$\hat{\theta}_{GLS} = \left(\Phi^{\top} V^{-1} \Phi \right)^{-1} \Phi^{\top} V^{-1} \mathbf{y}, \quad (3.34)$$

where V is the error correlation matrix. GLS is implemented in an iterative manner, where parameter convergence is monitored and V is updated.

In MSD, where a large number of possible model terms are considered to represent a system dynamics, obtaining parameter estimates of the superset using LS could be misleading, as overfitting is highly viable. The probability of poor conditioning increases with the matrix dimension. Therefore regularised least squares (RLS) is employed in cases like this, where a regularisation parameter α which is

a positive real number regularises the influence of the data set on the parameter estimates. The RLS can be defined as

$$\hat{\boldsymbol{\theta}}_{RLS} = (\alpha I + \Phi^T \Phi)^{-1} \Phi^T \mathbf{y}, \quad (3.35)$$

where I is an identity matrix.

Nonlinear methods

If the gradient of a minimising cost function is nonlinear-in-the parameters $\boldsymbol{\theta}$, eqn(3.21) is no longer valid, therefore a nonlinear optimisation technique is needed to search for the optimal parameters,

$$\mathbf{g}_n = \frac{dJ(\boldsymbol{\theta})}{d\boldsymbol{\theta}}, \quad (3.36)$$

where \mathbf{g}_n is the gradient vector of the cost function $J(\boldsymbol{\theta})$.

There are various different types of nonlinear optimization techniques classified as either local or global search. In this thesis, local search is only of interest. The commonly used gradient-based methods shall be reviewed. Regarding gradient-based methods, the gradient \mathbf{g}_n is assumed to be estimated by analytical calculations or approximated by finite difference techniques. The gradient-based methods are generally captured by the following equations (Nelles, 2001)

$$\boldsymbol{\theta}^j = \boldsymbol{\theta}^{j-1} - \eta^{j-1} \mathbf{p}_n^{j-1}, \quad (3.37)$$

$$\mathbf{p}_n^{j-1} = R_n^{j-1} \mathbf{g}_n^{j-1}, \quad (3.38)$$

where η is the step size with direction vector \mathbf{p}_n , which is evaluated based on the gradient direction vector \mathbf{g}_n , the direction scaling matrix R_n and j is the iteration number.

The steepest decent (SD) is the simplest form of the gradient-based methods, which is obtained by replacing the direction scaling matrix R_n with an identity matrix I . SD is easy to implement as the minimisation of the cost function is a first order derivative and requires only linear computation. However, when little prior knowledge of the parameter $\boldsymbol{\theta}$ is known, SD would have a slow convergence and is typically not applied.

Techniques involving second order derivative for the minimisation of the cost

function are classed under the Newton's method (NM). The direction scaling matrix R_n is chosen to be the inverse of the Hessian \mathbb{H}^{-1} . For robustness issues the value of η is mostly adjusted by the line search. \mathbb{H} is forced to be a positive definite during initialisation in order to assure decrease in the cost function. NM is more computational demanding than the SD as it calculates the inverse of the Hessian matrix, but it is much faster in terms of convergence.

Separable least squares

Separable least squares (SLS) was introduced in 1973 (Golub and Pereyra, 1973), which is regarded as a method to estimate the model parameters, where it separates the parameters into linear and nonlinear sets. The advantages of the SLS method is that it typically converges in fewer iterations, has improved numerical conditioning and requires initialisation of fewer parameters in comparison to the full nonlinear optimisation problem (Bruls et al., 1999, Golub and Pereyra, 2003). The nonlinear parameters are initialised at the beginning and the linear parameters are estimated using linear methods thereby optimising it with respect to the nonlinear parameters. The nonlinear parameters are then updated using nonlinear methods whose cost function is only based on the nonlinear parameters. This is repeated until parameter convergence is achieved.

3.4 Model structure detection

A subtle distinction is taken in this thesis between model structure detection (MSD) and model selection (MS). MSD involves selecting a small number of model terms (regressors) from a superset of possible candidate model terms (regressors) to represent a system. Whereas, MS is the scoring of competing models, in order to evaluate the best model to describe the observed data and also general enough to predict other experimental datasets. In nonlinear systems, the number of possible candidate model terms available are usually very large. Therefore a rigorous and efficient identification of a parsimonious representation for a nonlinear system is crucial. As stated earlier, the models dealt with in this thesis are both linear and nonlinear -in-the-parameters, therefore a brief review on MSD for both linear and nonlinear regression models shall be discussed. It should be noted that an overview on linear regression techniques can be found in (Draper et al., 1966, Montgomery et al., 2012).

3.4.1 Linear regression

Theoretically, the most optimal MSD is the exhaustive search. It involves the using of a criterion to evaluate and select the best model formulated from a set consisting of models which are built from all possible combinations of model terms. This is closely related to MS which is discussed in the next section. This method is implemented when the number of combinations of model terms are very few. However in most cases, the number of models to be examined are very large and practically not efficient due to very high computation. Forward selection and backward elimination are two other methods in MSD. Forward selection involves adding model terms incrementally to a model (which was initially empty) based on their contribution in describing the system's dynamics. Backward elimination involves eliminating model terms from a model (which initially consisted of all possible candidate model terms), which is seen to be spurious in describing the system's dynamics. More details about forward selection and backward elimination can be seen in the review papers mentioned above.

A more effective method is the step-wise regression, which involves an iteration between adding a significant model term and removing a redundant model term (that is already included in the model) to and from a model by forward selection and backward elimination respectively. The redundant model terms arise due to the contribution provided by the newly added model terms. In (Billings and Voon, 1986), a step-wise method is implemented.

A common regression method, implemented for the MSD of nonlinear models is the orthogonal least squares (OLS), which selects model terms based on their contribution to the maximisation of an error reduction ratio (ERR) (Korenberg et al., 1988). The OLS robustly estimates the parameters of the models, as the usual solution to the ordinary LS can be inaccurate due to the need to compute the inverse of the information matrix ($\Phi^T \Phi$), which is often ill-conditioned. A modified method of the OLS-ERR, which is known as the forward regression orthogonal (FRO) was developed in Chen et al. (1989). The FRO circumvents testing the excessive number of all possible combinations of model terms, whereby it allows a methodical regressor and thus an efficient MSD. FRO and other related techniques have been developed and modified over the last 30 years, into a comprehensive and versatile framework for the estimation and validation of the NARMAX models (NARX model is a subclass of the NARMAX model) (Billings and Aguirre, 1995, Billings et al., 1988, Chen et al., 1989, Guo and Billings, 2007, Mao and Billings, 1997).

Most methods applied today are variants of the classical FRO method (Billings et al., 1988, Korenberg et al., 1988), and in this thesis FRO shall be used as a benchmark due to the wide body of literature available supporting the method. In implementing the FRO for the MSD of the NARX model, the process is a linear regression method. Whereas, when implemented for the NARMAX model the process becomes a pseudo-linear method, reason because the residual error is not known initially, but eventually computed in an iterative manner. It can also be stated that, the NARMAX model is linear-in-the-parameters, if the residual error is known. The FRO procedure for the MSD of the NARX model shall only be discussed (the NARMAX model is not implemented in this thesis).

Before describing the FRO procedure, orthogonalisation of the regression matrix Φ and the computation of the ERR shall be reviewed, as they are crucial steps in the FRO algorithm. Orthogonal decomposition is done by first partitioning the linear regression model (eqn(3.21) in matrix form)

$$\mathbf{y} = \Phi\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad (3.39)$$

$$\mathbf{y} = \Phi\mathcal{A}^{-1}\mathcal{A}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad (3.40)$$

$$\mathbf{y} = W\mathbf{g} + \boldsymbol{\epsilon}, \quad (3.41)$$

where $W = [\mathbf{w}_1, \dots, \mathbf{w}_{N_\theta}]$ is a $(N_y \times N_\theta)$ orthogonal regression matrix with orthogonal columns, \mathcal{A} is a $(N_\theta \times N_\theta)$ upper triangular matrix, $\boldsymbol{\epsilon}$ is the model residual error vector, \mathbf{g} is the corresponding parameter vector to be estimated and where

$$W = \Phi\mathcal{A}^{-1}, \quad (3.42)$$

$$\mathbf{g} = \mathcal{A}\boldsymbol{\theta}. \quad (3.43)$$

The QR decomposition of $\Phi = W\mathcal{A}$ is normally achieved by using modified Gram-Schmidt (MGS) algorithm (Chen et al., 1989). The columns in W are uncoupled, therefore their corresponding parameters in \mathbf{g} are also uncoupled. This allows one to evaluate the individual contribution of each model term (regressor) in W towards minimising the distance between the observed output and the one-step-ahead prediction.

The FRO procedure iteratively compares and ranks model terms by their measured significance. A model term's (regressor's) significance is measured by the contribution of the model term to the observed output based on the one-step ahead

prediction using ERR,

$$ERR_j = \frac{g_j^2 \mathbf{w}_j^\top \mathbf{w}_j}{\mathbf{y}^\top \mathbf{y}}, \quad (3.44)$$

where the one-step-ahead prediction of \mathbf{y} is given by $g_j \mathbf{w}_j$ and $j = 1, \dots, N_\theta$.

The FRO procedure for the MSD of the NARX model can be summarised as follows:

1. Compute the ERR value of each model term (regressor) and identify the model term (regressor) with the largest ERR value and move it to the first column of the regression matrix.
2. Orthogonalise the remaining model terms (regressors) with respect to the selected model terms (regressors) and then compute their ERR values. Choose the model term (regressor) with the largest ERR value and move it accordingly.
3. Perform step 2 until some threshold of ERR value is reached.

3.4.2 Nonlinear regression

If the residual error of a model appears in the regression matrix, then the MSD becomes a nonlinear regression process. Here, two methods are reviewed. Firstly, the work presented in Piroddi and Spinelli (2003), where a NOE model structure is set to be identified. The LS was implemented as the parameter estimation technique, which eventually makes the approach inconsistent in the model estimation. However, the focus will be on the corresponding MSD criterion, which is similar to the ERR. It differs slightly because a model term's significance is measured by the contribution of the model term to the observed output based on the simulation prediction rather than the one-step-ahead prediction. The simulation approach reported an improved model term selection under some restrictive conditions such as non-persistently exciting signals and fast sampling (Billings, 2013).

Secondly, the method to be reviewed is the bootstrap MSD developed in Kukreja et al. (2004). It should be noted that the model structure assumed here is the NARMAX model, however, the method can be easily expanded to other models that are nonlinear-in-the-parameters (such as the NOE model). The bootstrap MSD naturally generates the parameter statistics which is used for model term selection. It proceeds as: (i) compute the parameter estimates and the residual errors, (ii) generate the new bootstrap dataset by sampling the residuals with replacement, (iii)

form a new measurement using the predicted output and bootstrap dataset, (iv) estimate the corresponding parameters for the new measurement data, (v) build the parameter statistics and (vi) remove model terms whose parameter estimates cannot be distinguished statistically from zero.

3.5 Model selection

There could be several competing models which could be able to fit the observed data \mathbf{y} well enough. However, a scoring criterion is needed to evaluate the best model which fits the observed data best and also good enough to predict other experimental datasets. This process is called model selection (MS). Information criteria are mostly used in MS, which measures a trade-off between model performance (prediction power) and model complexity (number of model terms) (Ljung, 1999). The two most commonly used information criteria are Akaike's information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978), which can be represented as

$$AIC = \log\left(\left(1 + \frac{2N_\theta}{N_y}\right) \times \frac{1}{N_y} \sum_{k=0}^{N_y-1} e_k^2\right) \text{ and} \quad (3.45)$$

$$BIC = \log\left(\left(1 + \frac{N_\theta \log(N_y)}{N_y}\right) \times \frac{1}{N_y} \sum_{k=0}^{N_y-1} e_k^2\right) \quad (3.46)$$

respectively. The BIC criterion provides extra penalty to model complexity compared to the AIC criterion. The model with the smallest information criterion value is selected.

Another well applied MS procedure is the Bayes factor (an alternative to the classical hypothesis testing), which is derived under the Bayesian framework (Kass and Raftery, 1995). A benefit of the Bayes factor is, it provides the evidence of a model for or against a hypothesis which in this case is a competing model (check Table 3.2). The Bayes factor for comparing evidence supporting two models \mathcal{M}_i and \mathcal{M}_j is

$$B_f(i, j) = \frac{p(\mathcal{M}_i|\mathbf{y})/p(\mathcal{M}_j|\mathbf{y})}{p(\mathcal{M}_i)/p(\mathcal{M}_j)}. \quad (3.47)$$

Table 3.2: Interpretation of the Bayes factor

$B_f(i, j)$	Evidence against \mathcal{M}_j and in favour of \mathcal{M}_i
1-3	very weak
3-20	positive
20-150	strong
>150	very strong

3.6 Model validation

When a model has been estimated to represent a nonlinear system, its structure and parameters need to be tested to be correct. Prediction accuracy is an intuitive way of validating a model, however, one-step-ahead prediction do not account for the accumulation of prediction errors, therefore other prediction methods are needed to validate a model (Van et al., 1994).

A model is said to be general enough, if it can predict not only the observed data which is used for the estimation process, but also unseen experimental datasets. Therefore to test for a model's generalisation performance, the experimental dataset including all available data is usually split into two sets, the training data which is used for the estimation process and the test data which is used for the final assessment of the model estimated. This process is called cross-validation. For a more reliable cross-validation process, the simulation prediction is used, where the mean sum of squared error (MSSE) is computed to assess the model performance (using eqn(3.23)),

$$MSSE = \frac{1}{N_y} \sum_{k=0}^{N_y-1} (y_k - \hat{y}_k)^2. \quad (3.48)$$

3.7 Continuous-time system identification

As discussed earlier in section 3.2, the nonlinear black-box models implemented in this thesis are in continuous-time (CT). A brief review about DT system identification was presented above, as it is much more established in comparison to CT system identification. Most dynamical systems encountered in practise are both continuous in time and nonlinear. Naturally in this present "digital world" with cheap computing power and digital electronics, system identification is dominated by DT techniques (Unbehauen and Rao, 1990). Nevertheless, the relevance and development of CT system identification especially for linear systems, have

been increasingly recognised and grown in last 20 years. Distinctly, the advantages of CT system identification towards the work presented in this thesis shall be outlined (Garnier and Wang, 2008):

- Easy interpretability and better physical understanding of parameters permits system properties to be transparent in the CT form. As this thesis involves multidisciplinary work, the results and analyses are to be shared amongst biologist and engineers, therefore the mentioned properties are crucial.
- Biological systems are very complex and intricate. CT models have compact representation (in comparison to DT) with good predictive power, which aids rapid and effective design procedures.
- The CT models are independent of the sampling period therefore permit identification of irregular sampled data. Most experimental data obtained from biological systems are irregularly sampled.
- CT models are more suitable for fast sampling and less prone to ill-conditioning.

The DT nonlinear black-box models shown in section 3.2.2 can be formulated into equivalent CT models. The CT-NARX and CT-NOE models are of key interest. Details for the apparent choice of the model structures can be seen in the respective chapters it is used in. The CT-NARX model can be represented as

$$y^{n_i}(t) = f\left(y(t), \dots, y^{n_i-1}(t), u(t), \dots, u^{n_i-1}(t)\right) + e(t), \quad (3.49)$$

where $y(t) \in \mathbb{R}$, $u(t) \in \mathbb{R}$ and $e(t)$ are the equivalent CT output, input and noise signal at time t . The derivative order is n_i and the mapping function $f(\cdot)$ consist of output and input signal derivatives. The CT-NOE model can be represented as

$$z^{n_i}(t) = f\left(z(t), \dots, z^{n_i-1}(t), u(t), \dots, u^{n_i-1}(t)\right), \quad (3.50)$$

$$y(t) = z(t) + e(t), \quad (3.51)$$

where $z(t) \in \mathbb{R}$ is the undisturbed CT output signal at time t .

3.7.1 Model estimation for continuous-time system identification

As stated earlier in section 3.3 and 3.4, model estimation is a dual problem of parameter estimation and MSD. The parameter estimation methods reviewed in

section 3.3 can be equally applied to CT models (Garnier and Wang, 2008, Garnier et al., 2003). MSD methods in CT system identification include direct and indirect methods. Direct methods identify CT models directly from the observed data, whereas, indirect methods initially identify DT models and then map it into the CT domain (Billings, 2013, Rao and Unbehauen, 2006, Unbehauen and Rao, 1990).

MSD for nonlinear CT models in system identification is not widely researched in comparison to the CT linear domain. The majority of the methods available for nonlinear CT domain are direct methods. They are developed based on DT-MSD methods permitting regression-based techniques. However, this was only possible because the estimation of output and input signal derivatives was plausible. The estimation of signal derivatives is a challenging task, since the observed output is typically corrupted by high frequency measurement noise and approximating the derivatives from directly differencing the observed signal amplifies this noise.

In Tsang and Billings (1994), signal derivatives are recovered by using delayed state-variable filters. It involves passing the input and output signals through multiple pre-filters with user specified bandpass. In Anderson and Kadiramanathan (2007), the delta-operator is used to map DT signals to the delta domain, which in effect retrieves the signal derivatives. The method developed in Coca and Billings (1999), will be used as a benchmark in this thesis due to its robustness in estimating signal derivatives compared to the other two methods mentioned earlier. It employs the FRO as the MSD method and it shall be termed the derivative continuous-time method (dCTM).

3.7.2 Signal derivative estimation using dCTM

The dCTM recovers signal derivatives using Kalman smoothing. Kalman smoothing is implemented on a state-space model, which is formulated based on the Taylor-series expansion of the observed signal. The Taylor-series expansion exploits the regularity of the solutions of differential equations, which was initially proposed by Fioretti and Jetto (1989). The observed signal's derivatives can be represented in the DT state space model as

$$\mathbf{x}_{k+1} = A\mathbf{x}_k + \mathbf{v}_k^1, \quad (3.52)$$

$$y_k = C\mathbf{x}_k + v_k^2, \quad (3.53)$$

where $C = (1, 0, \dots, 0)$ is the measurement matrix, $\mathbf{x}_k = (y_k, \dot{y}_k, \dots, y_k^D) \in \mathbb{R}^{n_x}$ is the state vector at sample time k that contains the vector of observed output and its derivatives up to order D , $n_x = D + 1$, $v_k^1 \sim \mathcal{N}(0, Q)$ and $v_k^2 \sim \mathcal{N}(0, R)$ are independent zero mean Gaussian white noise signals, and the state transition matrix A is described using rows that are based on the Taylor series expansion of the signal,

$$A = \begin{bmatrix} 1 & T & \frac{T^2}{2!} & \dots & \frac{T^D}{D!} \\ 0 & 1 & T & \dots & \frac{T^{D-1}}{(D-1)!} \\ \vdots & & & & \vdots \\ 0 & 0 & \dots & \dots & 1 \end{bmatrix}. \quad (3.54)$$

The elements of the state noise covariance matrix Q are given by

$$q_{ij} = \frac{\sigma_w^2 T^{2D+3-(i+j)}}{(D+1-i)!(D+1-j)!(2D+3-(i+j))'}, \quad (3.55)$$

where σ_w is a tuning parameter describing the power of the state noise. In order to obtain the derivatives in the state vector, Kalman filter and Rauch-Tung-Striebel smoother was used in this thesis.

3.8 Generalised frequency response functions

The analysis of nonlinear system in the frequency domain can provide important insights into a system's nonlinearity and physical behaviour. In comparison to the linear systems, research on frequency domain methods for nonlinear systems has received very limited research (Billings, 2013). Generalised frequency response functions (GFRFs) are higher-order functions that are multi-dimensional, which are used in representing nonlinear systems in the frequency domain.

The spectral analysis of nonlinear systems can be described by the Volterra series (Volterra, 2005). The general form of the Volterra series representation of a CT nonlinear system can be expressed as

$$y(t) = \sum_{n_f=1}^{N_f} y_{n_f}(t), \quad (3.56)$$

where

$$y_{n_f}(t) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_{n_f}(\tau_1, \dots, \tau_{n_f}) \prod_i^{n_f} u(t - \tau_i) d\tau_i, \quad (3.57)$$

where $h_{n_f}(\tau_1, \dots, \tau_{n_f})$ is the system's n_f th order Volterra kernel or impulse response and N_f is the maximum order of system nonlinearities which is finite for a wide class of nonlinear systems and input excitations according to the analysis by (Boyd and Chua, 1985).

In George (1959), the n_f th order GFRF is defined by the Fourier transform of the n_f th order Volterra kernel $h_{n_f}(\tau_1, \dots, \tau_{n_f})$,

$$H_{n_f}(j\omega_1, \dots, j\omega_{n_f}) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_{n_f}(\tau_1, \dots, \tau_{n_f}) e^{-j(\omega_1\tau_1 + \dots + \omega_{n_f}\tau_{n_f})} d\tau_1 \dots d\tau_{n_f} \quad (3.58)$$

where $n_f = 1, 2, \dots$

The first order GFRF $H_1(j\omega)$ is used to explain the linear properties, while the nonlinear GFRFs $H_{n_f}(j\omega_1, \dots, j\omega_{n_f})$ for $n_f > 1$ describes the nonlinear phenomena. The nonlinear phenomena that can be explained are harmonics, gain compression and expansion, desensitisation and intermodulation (Billings et al., 1990).

The estimation and computation of the GFRFs can be approached in the parametric model-based method. The parametric model-based method is applied to nonlinear models that are already been derived, by mapping it analytically into the frequency domain (Billings and Tsang, 1989a,b). This is done by discarding the noise model and only taking the deterministic part of the parametric model into account. An advantage of this method is that the relationship between the time and frequency domain behaviours is not lost, as each model term's influence on each GFRF can be clearly seen (Billings, 2013). This helps physical interpretation and design procedures, as certain nonlinear properties are designed in the frequency domain but implemented in the time domain.

3.8.1 Probing method

The GFRFs of a parametric nonlinear model can be derived analytically using the probing method. Using a combination of N_f exponentials,

$$u(t) = \sum_{i=1}^{N_f} e^{j\omega_i t}, \quad (3.59)$$

to excite a nonlinear systems represented by Volterra series, the n_f th order output response can be written as,

$$y_{n_f}(t) = \sum_{i_1=1}^{N_f} \dots \sum_{i_{n_f}=1}^{N_f} H_{n_f}(j\omega_{i_1}, \dots, j\omega_{i_{n_f}}) e^{j(\omega_{i_1} + \dots + \omega_{i_{n_f}})t}. \quad (3.60)$$

A recursive algorithm is developed to help automate the process of estimating and computing GFRFs (Jones and Billings, 1989).

3.9 Summary

Nonlinear black-box models from system identification has been successfully applied for data-driven modelling and analysing of engineering, financial, medical and environmental systems. However, despite the versatility of the approach, it is essentially novel in the field of synthetic biology. In this chapter, a brief introduction to system identification have been reviewed. The main steps in identifying a nonlinear model involves parameter estimation, model structure detection, model selection and model validation.

In relation to the work presented in this thesis, the need to identify data-driven nonlinear models in continuous-time is emphasised. Which provides a parsimonious representation of the biological systems and a good generalisation performance. The concept of generalised frequency response transfer functions was also introduced as a powerful tool for analysing estimated system dynamics in the frequency domain. To summarise, this chapter has provided ample evidence that nonlinear black-box models and analysis methodologies provide the robust choice for characterisation of genetic parts in biological systems. This will be further demonstrated in later chapters.

Chapter 4

Modelling a transcriptional regulation

4.1 Introduction

A key challenge in synthetic biology is the development of effective methodologies for characterisation of genetic parts in biological systems, in a form suitable for dynamic analysis and design. In the earlier chapter, nonlinear system identification was introduced, that provides a comprehensive package of tools which would aid in overcoming this challenge. In this chapter, it is demonstrated by implementing a data-driven nonlinear dynamic modelling framework that is popular in the field of control engineering, but is novel to the field of synthetic biology. The framework is applied to identify a "population-level" nonlinear autoregressive model with exogenous input (NARX) to represent the transcriptional regulatory genetic part BBa_T9002 that is obtained from the registry of standard biological parts (RSBP), which serves as a case study. The developed "population-level" NARX model exhibits accurate representation of the system dynamics, a structure that is parsimonious and consistent across different cell populations. Dynamic and static biochemical models (from the literature) which are derived from simple enzymatic reaction scheme (ERS), are used as benchmarking comparisons. The identified data-driven model is shown to be comparably simple, but exhibits much greater prediction accuracy on the experimental data.

With the advances in the field of synthetic biology as discussed in Chapter 2, there are yet a number of obstacles to overcome before biological systems can be transformed from laboratory prototypes to industrial products that would be applicable to practical problems (Arkin, 2008, Kwok, 2010). Embracing the engineering con-

cept of model-based design is crucial, as challenges in characterisation and design can be largely reduced, in particular, to underpin the future success of top-down biological synthesis using off-the-shelf genetic parts (obtained from repositories) (Andrianantoandro et al., 2006, Endy, 2005). Here, the recent results in static input-output characterisation (Canton et al., 2008), is extended by developing a data-driven framework for describing the dynamic properties (equally static) of genetic parts in biological systems. The models derived using this framework is shown to be useful and therefore recommended to be specified in datasheets associated with genetic parts, with the purpose of aiding in the control design and synthesis of larger systems.

The data-driven framework presented here differs with the modelling approaches commonly applied in the field of synthetic biology. Gene expressions (English et al., 2005, Gertz et al., 2008, Houston and Kenworthy, 2000, Kim et al., 2006) are mostly modelled using simplifications of ERS, *e.g.* the Michaelis-Menten (MM) and Hill equations (Cornish-Bowden, 2013). These models are relatively simple to implement, but pose limitations, and retain a fixed model structure regardless of the complexity of the system under consideration. Also, network-based dynamic gene expression models that have been explored are in the form of many coupled ordinary (Covert et al., 2008, Karlebach and Shamir, 2008, Tian and Burrage, 2006) or stochastic (Wilkinson, 2009) differential equations (ODEs and SDEs respectively). As the genetic regulatory network grows larger, the number of equations needed for the representation of the biological system increases, leading to an explosion in model complexity. The models in the literature are mostly either incapable or intractable for dynamic systems analysis and design.

In this chapter, a data-driven framework to identify a "population-level" NARX model for a transcriptional regulatory genetic part is undertaken. A framework is suggested to provide useful models which are able to overcome the typical problems of existing models used in describing genetic parts in biological systems, which are overly complex, unwieldy and of unknown structure. The NARX model, a subclass of the more general NARMAX model which provides a general nonlinear dynamic system description. The NARX model identified is able to have a compact representation that: (i) represents both the static and dynamic properties, (ii) represents both the deterministic and stochastic processes of the system via noise term and (iii) is flexible and adaptable in its structure and parameters according to the experimental data (since it is data-driven).

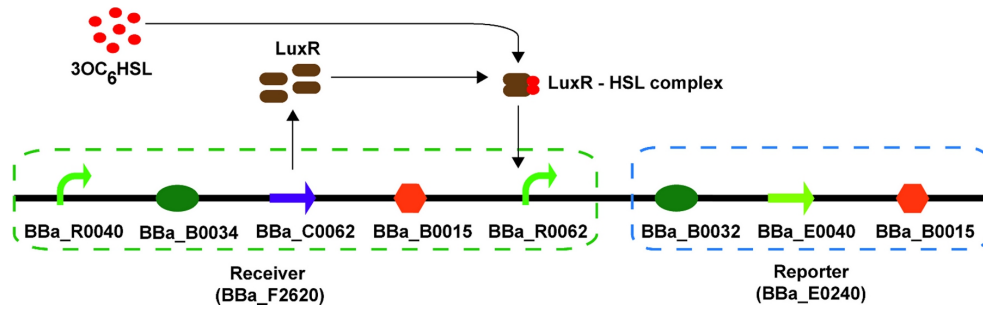


Figure 4.1: Pictorial description of the BBa_T9002 system with input and output of 3OC₆HSL and GFP expression respectively.

4.2 The BBa_T9002 system and its experimental data

The genetic part BBa_T9002, is a quorum sensing receiver-reporter composite system shown in Figure (4.1): The receiver (BBa_F2620) and reporter (BBa_E0240). The label - BBa for the genetic parts are identity codes assigned to them in the RSBP.

Cells, in the absence of tetracycline and TetR, can constitutively express and up-regulate LuxR on addition of 3-ox-ohexanoyl-L-homoserine lactone (3OC₆HSL). In a ratio of 2 : 2, the LuxR protein forms a complex with the signalling molecule 3OC₆HSL, which activates the LuxR regulated promoter BBa_R0062 producing the receiver's (BBa_F2620) output, quantified as polymerases per second (PoPS). The activation of the LuxR regulated promoter subsequently prompts the expression of green fluorescence protein (GFP), which serves as the output for the BBa_T9002 system.

The advantage of using the BBa_T9002 system as a case study is:

- It is effectively a single transcriptional regulatory genetic part (transcription-translation system). Single transcriptional regulation systems are one of the most simplistic genetic functional modules and are frequently used as the foundational modules to design higher-order genetic parts.
- It has been well studied and characterised using alternative modelling techniques (Canton et al., 2008) and the experimental data is available online, facilitating further comparison and investigation.

The experimental data used in this chapter, which involves the dynamic response of the genetic part BBa_T9002 was collected by Canton et al. (2008). The genetic

part, BBa_T9002, was transformed into the Escherichia coli (*E. coli*) wild-type strain MG1655, which was streaked onto plates to obtain colonies for experimentation. The cultures of the colonies were grown in M9 minimal media for 15 hours at 37°C with shaking at 70 revolution per second (*rpm*). The absorbance (cell growth) and fluorescence (GFP) measurements were obtained repeatedly using the Wallac Victor3 multi-well fluorimeter.

The experimental data can be obtained from the RSBP website for part BBa_F2620 (http://partsregistry.org/Part:BBa_F2620). The experimental data consisted of time-series recording of GFP expression, which was observed over approximately 180 minutes (77 time steps, sampled at intervals of approximately 141 seconds) of BBa_T9002 over 8 different 3OC₆HSL input concentrations: 0, 1e-10, 1e-9, 1e-8, 1e-7, 1e-6, 1e-5, 1e-4 molar (*M*). In this investigation, systems modelling was performed up to the point of quasi steady state behaviour (see below), which was defined as the peak of the rate of GFP expression. This truncation resulted in time-series data that were approximately 150 minutes in length (approximately 60 time samples; the range varied between 50-70 time samples among different experiments). Here, experimental data were analysed from 3 colonies of BBa_T9002 (out of a total of 9 observed by Canton et al. (2008)). There were 3 replicates for each colony, resulting in 9 experiments in total that are analysed. Experiments 1-3, 4-6 and 7-9 were from colonies 1, 2 and 3 respectively.

4.3 Data pre-processing - signal derivative estimation

All models in this chapter are directly identified in continuous-time (CT). The derivatives of the GFP expression was estimated using the Kalman smoothing algorithm as shown in section 3.7.2 - under dCTM, which led to derivative estimates that were relatively noise-free compared to directly differenced signals (Figure 4.2).

The GFP expression signal vector shall be denoted as $\tilde{\mathbf{y}}$, therefore eqn(3.53) can be re-written as

$$\tilde{\mathbf{y}}_k = \mathbf{C}\mathbf{x}_k + \nu_k^2, \quad (4.1)$$

where $\mathbf{x}_k = (\tilde{\mathbf{y}}_k, \dot{\tilde{\mathbf{y}}}_k, \dots, \tilde{\mathbf{y}}_k^D)$. Kalman filter and Rauch-Tung-Striebel smoother was used in implementation.

Algorithm 4.1 Kalman filter**Initialise:** $P_{0|0}$ and $\mathbf{x}_{0|0}$ **for** $k = 1 : N_y$

Predictor equations

Predict state: $\hat{\mathbf{x}}_{k|k-1} = A\hat{\mathbf{x}}_{k-1|k-1}$ Predict state covariance: $P_{k|k-1} = AP_{k-1|k-1}A^\top + Q$

Corrector equations

Calculate Kalman gain: $K_k = P_{k|k-1}C^\top (CP_{k|k-1}C^\top + R)^{-1}$ Correct state: $\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + K_k(\tilde{y}_k - C\hat{\mathbf{x}}_{k|k-1})$ Correct state covariance: $P_{k|k} = P_{k|k-1} - K_kCP_{k|k-1}$ **end for****4.3.1 Kalman filter**

In order to estimate the signal derivatives, the Kalman filter (KF) recursions can first be used to retrieve $\hat{\mathbf{x}}_{k|k}$ for $k = 1 : N_y$, the estimate of the state \mathbf{x} at time k based on the measurements up to time k . In 1960 (Kalman and others., 1960), Kalman filtering was introduced which could be used in predicting and estimating the present, past and future states of a linear system given its model and observation data. The KF is a recursive estimator, and it is an optimal estimator for linear systems with Gaussian noise. The equations of KF fall into two groups, predictor equations and corrector equations (Welch and Bishop, 1995). The predictor equations project the current state estimates ahead of time (using the observation at current time and the state estimate at previous time step) and the corrector equations adjust the projected current state estimate using the actual observation at that time.

If the Kalman gain, a priori state covariance and a posteriori state covariance are K_k , $P_{k-1|k-1}$ and $P_{k|k-1}$, then the KF algorithm can be mathematically described as shown in Algorithm 4.1.

4.3.2 Rauch-Tung-Striebel smoother

The backward recursions to obtain the smoothed state $\hat{\mathbf{x}}_{k|N_y}$ can be implemented using the Rauch-Tung-Striebel smoother (RTSS). RTSS was developed in 1965 (Rauch et al., 1965), which can be used in conjunction with KF for smoothing. Smoothing estimates the current states using all the observation, which is normally applied after filtering in order to smooth out the filtered state estimates. KF and RTSS are implemented on a backward-forward framework, where KF is used for the forward pass and RTSS includes the backward pass. For a linear state space system with additive Gaussian noise, smoothed state covariance ma-

Algorithm 4.2 Rauch-Tung-Striebel smoother

Forward pass

Run: Algorithm 4.1Store filtered states $\hat{\mathbf{x}}_{k|k}$ and their corresponding covariances $P_{k|k}$ **Initialise:** $P_{N_y|N_y}$ and $\hat{\mathbf{x}}_{N_y|N_y}$ at $k = N_y$

Backward pass

for $k = N_y - 1 : 1$

$$J_k = P_{k|k} A^\top P_{k+1|k}^{-1}$$

$$\text{Smooth state: } \hat{\mathbf{x}}_{k|N_y} = \hat{\mathbf{x}}_{k|k} + J_k (\hat{\mathbf{x}}_{k+1|N_y} - A \hat{\mathbf{x}}_{k|k})$$

$$\text{Smooth covariance: } P_{k|N_y} = P_{k|k} + J_k (P_{k+1|N_y} - P_{k+1|k}) J_k^\top$$

end for

trix $P_{k|N_y}$, smoother gain J_k , the RTSS can be described by Algorithm 4.2.

In this investigation, the following values for these parameters were used: number of derivative terms, $D = 6$, initial state uncertainty $P_{0|0} = 100 \times I_{n_x}$, state dimension $n_x = D + 1 = 7$ and σ_w^2 was set to 10^{-5} .

4.4 Identification of dynamic and static biochemical models

Two different approaches to modelling the BBa_T9002 system was taken here: (i) dynamic and static biochemical models based on a well known description of an ERS and (ii) a data-driven CT-NARX model. The first approach (in this section) is shown to have certain drawbacks which the second approach (next section) is able to provide solutions to.

The work presented in Canton et al. (2008), is used as a foundation to this section and a benchmarking comparison to the data-driven CT-NARX model presented in this chapter. Canton et al. (2008), derived a Hill equation model for BBa_F2620 genetic part by indirectly using the experimental data collected for BBa_T9002 (the knowledge of the model for BBa_E0240 was assumed). In contrast here, the complete BBa_T9002 system is modelled.

4.4.1 Derivation of dynamic and static biochemical models

The Michealis-Menten and Hill equations are well known models used in describing enzymatic reactions and are widely applied in the synthetic biology literature (section 2.3.2). These models are derived by applying assumptions to nonlinear coupled ODEs used in describing the ERS, thereby simplifying them (see below

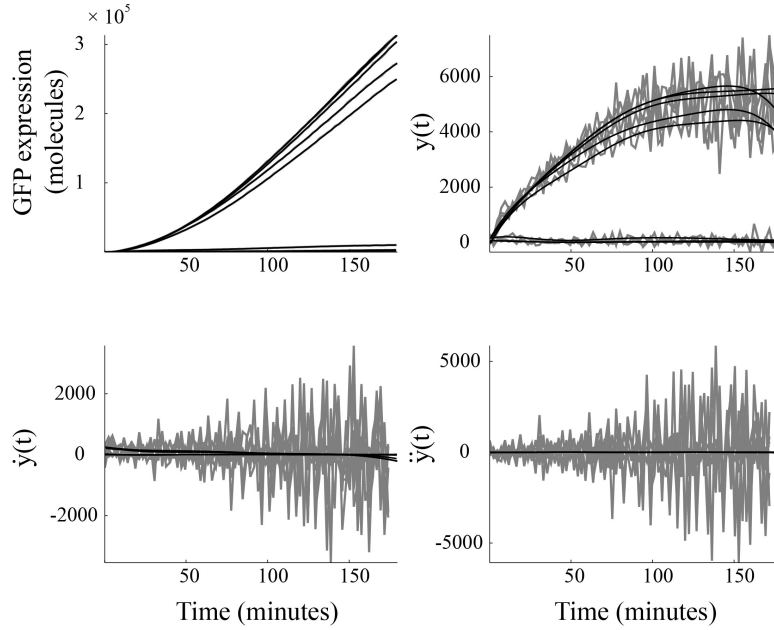


Figure 4.2: The GFP expression signal and its derivatives obtained from the RTS smoothing algorithm (black) in comparison to derivatives from numerically differencing the raw GFP expression signal (grey).

for derivation). A major drawback of the Michealis-Menten and Hill equations is that they do not provide a full dynamic description of the system, thereby not facilitating design procedures; only the static relationship between product derivative and substrate input is captured. In order to address this drawback, a model is derived which is used to describe the dynamics of the ERS associated with the Hill equation, which is denoted as the ERS model.

Considering the reaction scheme in section 2.3.2 and the nonlinear coupled ODEs used in describing it - eqn(2.5, 2.6, 2.7 and 2.8), the following simplified equations can be written in terms of substrate and product variables only, where: (i) $n = 1$ and (ii) substrate s and product p are 3OC₆HSL and GFP expression respectively, in this investigation,

$$\dot{s}(t) = -k_1 e_0(t) s(t) + \frac{k_{-1}}{k_2} \dot{p}(t) + \frac{k_1}{k_2} s(t) \dot{p}(t), \quad (4.2)$$

$$\ddot{p}(t) = k_2 k_1 e_0(t) s(t) - (k_{-1} + k_2) \dot{p}(t) - k_1 s(t) \dot{p}(t), \quad (4.3)$$

where e denotes enzyme, s substrate, n Hill coefficient, e_s enzyme-substrate complex, p product and where the reaction scheme has an associated total enzyme

concentration e_0 . The parameters k_1 , k_{-1} and k_2 defines the rate of reactions. The eqn(4.2) is obtained from rearranging eqn(2.8), so that $e_s(t) = \frac{\dot{p}(t)}{k_2}$, noting that $e(t) = e_0(t) - e_s(t)$ and substituting for $e_s(t)$ and $e(t)$ in eqn(2.5); eqn(4.3) is obtained from substituting eqn(2.7) and $e(t) = e_0(t) - e_s(t)$ into $\dot{p}(t) = k_2 e_s(t)$.

Michealis-Menten equation

The assumptions made in deriving the Michealis-Menten equation are: (i) the total concentration of enzyme $e_0(t)$ does not change over time, $e_0(t) = e_s(t) + e(t)$ and (ii) the rate of change of enzyme-substrate complex $e_s(t)$ is approximately zero which is referred to as quasi steady state, $\dot{e}_s(t) = 0$.

For the assumptions to hold, provided that the parameters are not time-varying, the states mentioned below are at steady state: (i) $e_s(t)$, $\dot{e}_s(t)$ is equal to zero, (ii) $e(t)$, in order to satisfy assumption (i) above, and (iii) $s(t)$, in order to satisfy eqn(2.7).

The Michealis-Menten equation is derived by first equating eqn(2.7) to zero and substituting, $e_{ss} = e_0(t) - e_{sss}$ (where subscript ss refers to steady state)

$$(k_{-1} + k_2 + k_1 s_{ss}) e_{sss} = k_1 e_0(t) s_{ss}, \quad (4.4)$$

$$e_{sss} = \frac{e_0(t) s_{ss}}{s_{ss} + k_m}, \quad (4.5)$$

where $k_m = \frac{k_{-1} + k_2}{k_1}$. Substituting eqn(4.5) into eqn(2.8), provides the Michealis-Menten equation (Cornish-Bowden, 2013)

$$\dot{p}(t) = k_2 e_{sss} = v_{max} \left(\frac{s_{ss}}{s_{ss} + k_m} \right), \quad (4.6)$$

where $v_{max} = k_2 e_0(t)$.

Enzymatic reaction scheme model and Hill equation

The steps carried out above can be extended to obtain an ERS model and the Hill equation for $n \neq 1$. The corresponding ODEs (in terms of substrate and product variables only) - the ERS model can be written as

$$\dot{s}(t) = a_1 s^n(t) + a_2 \dot{p}(t) + a_3 s^n(t) \dot{p}(t), \quad (4.7)$$

$$\ddot{p}(t) = b_1 s^n(t) + b_2 \dot{p}(t) + b_3 s^n(t) \dot{p}(t), \quad (4.8)$$

where $a_1 = -nk_1e_0(t)$, $a_2 = \frac{nk_{-1}}{k_2}$, $a_3 = \frac{nk_1}{k_2}$, $b_1 = k_2k_1e_0(t)$, $b_2 = -k_1 - k_2$ and $b_3 = -k_1$. For the assumptions mentioned above in deriving the Michealis-Menten equation, total enzyme concentration and the rate of change of enzyme-substrate complex to be constant and equal to zero respectively. The more often used Hill equation, which is a simplification of the full dynamic form described in eqn(4.7 and 4.8), can be derived

$$\dot{p}(t) = v_{max} \left(\frac{s_{ss}^n}{s_{ss}^n + k_p^n} \right), \quad (4.9)$$

where $k_m = k_p^n = \frac{k_{-1}+k_2}{k_1}$ and $v_{max} = k_2e_0(t)$.

Note that the Hill equation should provide a description of the reaction scheme that is consistent with the dynamic form presented above in eqn(4.7 and 4.8) for substrate signals that are near constant. Also note that in the case of steady state substrate, the LHS of eqn(4.7) is by definition equal to zero, *i.e.*, $\dot{s}(t) = 0$. Hence, any dynamic behaviour in the system must be predominantly described by eqn(4.8).

4.4.2 Parameter estimation of enzymatic reaction scheme model and Hill equation

Enzymatic reaction scheme model

The following prior information was used in estimating the parameters of the ERS model: (i) initial 3OC₆HSL level, $s(t = 0)$, and (ii) GFP expression over time, $p(t)$. The separable least squares SLS algorithm (Bruls et al., 1999, Golub and Pereyra, 2003) was implemented to estimate the model parameters. The SLS algorithm allowed the separation of the complete parameter set into into linear and nonlinear sets, $\zeta_1 = (b_1, b_2, b_3)$ and $\zeta_n = (a_1, a_2, a_3, n)$ respectively. This is advantageous because the parameter convergence is much faster, has improved numerical conditioning and fewer parameters are required to be initialised in comparison to the full nonlinear optimisation problem (Golub and Pereyra, 2003) (as discussed in section 3.3.2).

The optimisation cost function for M experimental signals corresponding to different input levels of 3OC₆HSL, with N_y samples per signal can be defined as

$$J = \frac{1}{MN_y} \|\mathbf{P} - \Gamma(\zeta_n)\zeta_l\|_2^2, \quad (4.10)$$

where $\mathbf{P} = [\mathbf{p}_1^\top \dots \mathbf{p}_M^\top]^\top$, $\mathbf{p}_j = [\dot{p}_j(1) \dots \dot{p}_j(N_y)]^\top$, $\Gamma(\zeta_n) = [\gamma_1(\zeta_n)^\top, \dots, \gamma_M(\zeta_n)^\top]^\top$ and

$$\gamma_j(\zeta_n) = \begin{bmatrix} \hat{s}_j^n(1) & \dot{p}_j(1) & \hat{s}_j^n(1)\dot{p}_j(1) \\ \vdots & \vdots & \vdots \\ \hat{s}_j^n(N_y) & \dot{p}_j(N_y) & \hat{s}_j^n(N_y)\dot{p}_j(N_y) \end{bmatrix}, \quad (4.11)$$

and $\hat{s}_j(\cdot)$ is obtained from simulation of eqn(4.7), given the initial condition of the substrate and the nonlinear parameters ζ_n . The linear parameters ζ_l can be expressed in terms of the nonlinear parameters ζ_n using the LS solution

$$\zeta_l = \Gamma(\zeta_n)^\dagger \mathbf{P}, \quad (4.12)$$

where \dagger denotes the pseudo-inverse, $\Gamma(\zeta_n)^\dagger = \left(\Gamma(\zeta_n)^\top \Gamma(\zeta_n)\right)^{-1} \Gamma(\zeta_n)^\top$. Substituting eqn(4.12) into eqn(4.10) leads to the reduced optimization problem, from which the linear parameters have been eliminated

$$\min_{\zeta_n} \frac{1}{MN_y} \|\mathbf{P} - \Gamma(\zeta_n)\Gamma(\zeta_n)^\dagger \mathbf{P}\|_2^2. \quad (4.13)$$

The quasi-Newton method was used to implement the nonlinear optimization (implemented using the MATLAB function `fminunc`). The initialisation of the nonlinear parameters were guided by using a grid search, in which the parameters ranges were $a_1, a_2, a_3 \in [-10, -9.9, \dots, 10]$, and $n \in [0, 0.5, \dots, 6]$. The MATLAB algorithm used for simulating the ERS model was based on the first order Euler approximation for computational simplicity (we verified on a subset of the data that use of higher order numerical integration methods did not alter the results).

Hill equation

The parameters of the Hill equation was estimated using nonlinear optimization technique - quasi-Newton method (implemented using the MATLAB function `fminunc`). The parameters v_{max} and k_p were initialised by trial and error as these two were the sensitive parameters that influenced convergence.

4.4.3 Results and discussion

In using the observations of the dynamic behaviour of the genetic part BBa_T9002, the ERS model was identified. By analysing the simulated results of the ERS model against the experimental data, the following was noted:

- The estimated level of the 3OC₆HSL signal remained constant from the time of induction to quasi steady state (Figure 4.3A); this may be due to the very high concentration of 3OC₆HSL molecules at the point of induction in comparison to the amount used up by the cells.
- The observed GFP expression were not well predicted by the ERS model (Figure 4.3C and D); the prediction error variance of the GFP signal was 24.3%. Due to the low prediction accuracy, the ERS model demonstrates that there is missing dynamics (model terms) in it or its model structure is not appropriate for describing the BBa_T9002 system.

The peak values of the rate of GFP expression (at 150th minute), indicated a sigmoidal function relationship with the substrate concentration (Figure 4.3B). This observation indicates cooperative binding is involved in the enzymatic reaction process (Cornish-Bowden, 2013). The Hill equation was used to fit the peak values as a function of 3OC₆HSL. The prediction of the steady state behaviour by the Hill equation was more accurate in comparison to the ERS model prediction as shown in Figure 4.3B.

The Hill equation is generally considered by the synthetic biology community to be a useful model for describing simple switching properties in biochemical processes (Canton et al., 2008, Tamsir et al., 2010, Wang et al., 2011). As shown above, the steady state GFP expression rate is more accurately predicted by the Hill equation than the ERS model (Figure 4.3B). However, the decay of the steady state behaviour at high levels of 3OC₆HSL was not captured by the Hill equation. This was a common feature across all experiments, which could be a result of toxicity to the cells (Canton et al., 2008).

The Hill equation and the ERS model are both derived from the same reaction scheme (see above). However, it is apparent from these results that the Hill equation provides a much improved model of the steady state process. The inconsistency between the dynamic and static model raises a question over the link between the Hill equation and the ERS model on which it is predicated, and hence the interpretability of the Hill equation parameters. One possible explanation

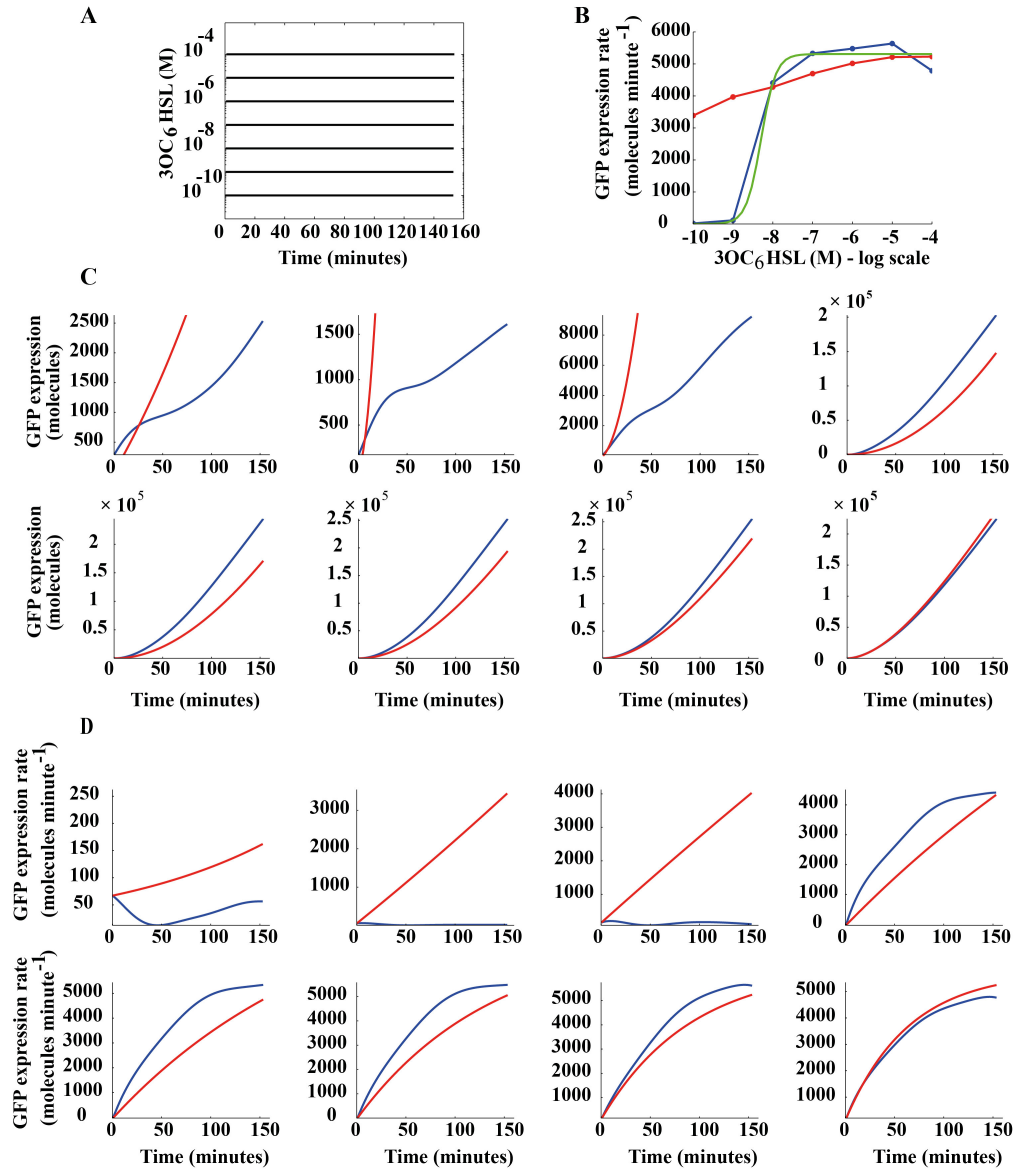


Figure 4.3: Biochemical model simulation for experiment 1: (A) The simulated model input signal $s(t)$, (B) Rate of change of GFP expression at the 150th minute (blue), the Hill equation prediction (green), and the ERS model prediction (red), (C) Comparison of GFP expression (blue) and the model prediction $p(t)$ (red) and (D) Rate of change of GFP expression (blue) and model prediction (red). Note that the response corresponding to the lowest input level $3OC_6HSL = 0$ in (A) has been omitted because of the log transformation, and the plots in (C) and (D) corresponds to responses to increasing input level from left to right - top to bottom.

could be for the improvement of the Hill equation over the ERS model is, that the sigmoidal form of the Hill equation is coincidentally well suited to describing the switching behaviour observed in the experimental data. It could also be suggested that the optimal ERS model parameters were unidentifiable or could not be obtained here due to the difficulties inherent in nonlinear parameter optimisation and the richness of the dynamics in the experimental data.

In the context of biosynthesis, model inaccuracies will be problematic: If a model fails to capture the key properties of a system, then errors will be imposed on the system design. This motivates the development of alternative modelling strategies that will solve these challenges.

4.5 Identification of data-driven dynamic model

The data-driven framework allows the identification of a nonlinear dynamic black-box models. For the identification of the BBa_T9002 system under this framework, the input and output signals are defined as 3OC₆HSL concentration and rate of change of GFP expression respectively. The rate of change of GFP expression was preferred as the output signal rather than GFP expression, because modelling a stable system is more desirable in a data-driven framework, and the initial growth in GFP is exponential (unstable), whereas its derivative is stable. The experimental dataset used here (collected by Canton et al. (2008)), does not have the input signal observed over time. Accordingly, the input signal is assumed to be constant which is equivalent to the initial concentration of 3OC₆HSL, for the relative short time-scale of the recording (from the time of induction to quasi steady state). Bioassay is a popular procedure to measure the concentration of 3OC₆HSL molecules over time. However, the measurement could turn out to be insignificant because the concentration levels are very low and the bioassay procedure itself is wasteful. Therefore the following hypothesis is suggested, the 3OC₆HSL molecules disintegrate and become available again after each complex formation with LuxR and transcriptional activation.

An advantage of the framework to identify a nonlinear black-box model is that the choice of model structure is data-driven. This is known as the model structure detection MSD problem, and there are a number of algorithms that can be used to automate the choices that determine model structure, *e.g.* dynamic order and basis function selection (Baldacchino et al., 2012, Chen et al., 1989, Wei et al., 2004). MSD is a powerful asset of the framework because it can highlight

"missing" dynamic, *i.e.*, terms that are absent from biochemically derived models, which are required to accurately describe the system. The nonlinear black-box model structure used here is the CT-NARX model, which has an advantage because it is linear-in-the-parameters. This is a useful feature, which facilitates rapid identification and comparison of many different proposed model structures. The CT-NARX model though predictive do not give a biophysical interpretation.

4.5.1 System description and physical insights

In each experiment, the response to each 3OC₆HSL induction approximately overlays each other when normalised to its response respectively (see Figure 4.4B). This is a feature which is well captured by cascade models (see section 3.2.3). Therefore, in this investigation the BBa_T9002 system is described by a single static and dynamic function by taking inspiration from the structure of a Hammerstein model.

In Canton et al. (2008), it is reported that the cell growth (OD₆₀₀) of each well in the 96 well plate was similar until they reach quasi steady state (which is only modelled here). Therefore, including a cell growth variable in the data-driven model would be uninformative in this case. This led to a model at a "population-level" (Figure 4.4A) which uses only the input and output data, capturing the GFP expression for different inductions of 3OC₆HSL.

The dynamic function in this case was chosen to be the CT-NARX model. Therefore the noise term $e(t)$ is assumed to capture the process noise - population heterogeneity which arises due to variability caused by intrinsic and extrinsic noise. A more complicated noise model was avoided, so that a simplified model is able to correctly identify the BBa_T9002 system which is the main focus. Measurement noise term was not included because smoothing of the experimental data was carried out.

4.5.2 CT-NARX model representation

Generally, the CT-NARX model is obtained in a data-driven framework from regularly sampled input and output signals. In this investigation, the input and output signals to the BBa_T9002 system are defined as $u(t) \in \mathbb{R}$ - 3OC₆HSL concentration and $y(t) \in \mathbb{R}$ - rate of change of GFP expression (due to its stable nature) respectively. In order to obtain the predicted GFP expression of the identified model, the output of the CT-NARX model - $y(t)$ is numerically integrated. The structure of a general CT-NARX model can be defined by (Coca and Billings, 1999)

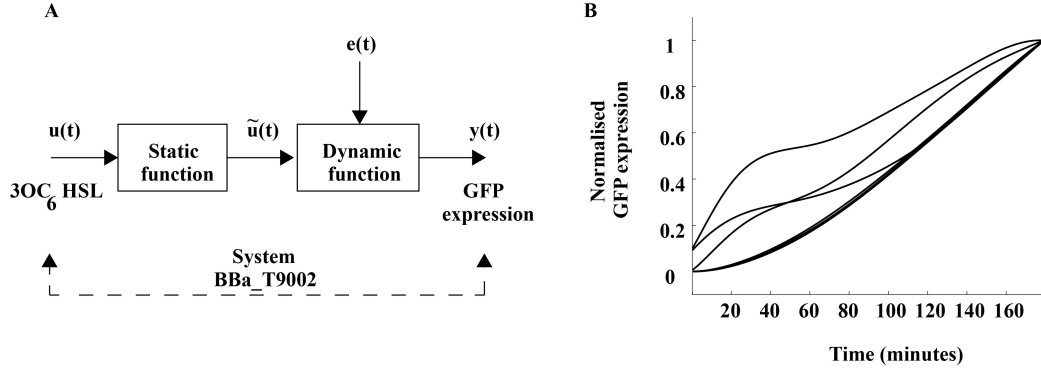


Figure 4.4: System representation and physical insight: (A) The model structure representation of the data-driven model where the dynamic function corresponds to the CT-NARX model and (B) The individually normalised GFP expression of each response in experiment 1. Normalisation is with respect to the final GFP expressions to remove the static gain effects.

$$y^{n_i}(t) = f(\mathfrak{S}(t)) + e(t), \quad (4.14)$$

$$\mathfrak{S}(t) = (y(t), \dots, y^{n_i-1}(t), u(t), \dots, u^{n_i-1}(t)), \quad (4.15)$$

where n_i is the differential order, $f(\mathfrak{S}(t))$ is some unknown nonlinear function and $\mathfrak{S}(t) \in \mathbb{R}^{2n_i}$ is the model input vector of system input and output derivatives. The function $f(\cdot)$ can be described using a basis function decomposition

$$y^{n_i}(t) = \sum_{j=1}^{N_\theta} \theta_j \phi_j(\mathfrak{S}(t)), \quad (4.16)$$

where $\phi_j(\cdot)$ is a basis function with associated real valued parameter $\theta_j \in \mathbb{R}$. In this investigation, polynomial basis functions of maximum order $q = 3$ was used and second order system dynamics, $n_i = 2$ was assumed.

Alterations to the general form of the CT-NARX model was imposed, in order to accommodate a specialised form for this investigation by only considering derivatives in the output signal and no cross-product terms between input and output signals. This specialised form was implemented because of the assumption that the input level of 3OC₆HSL was constant over the duration of each experiment, so the derivatives of the input signal were zero and the cross-product terms were unidentifiable. Constant variables can cause numerical ill-conditioning in regression equations due to the linear dependence that can arise between model

terms. Therefore, in this investigation where constant input is assumed, higher order polynomial input transformations and cross-product terms between input and output signals are not used. As mentioned above, there appeared to be a nonlinear gain variation associated with different input levels of 3OC₆HSL, which is described using separate input gain terms k_j , for $j = 1, \dots, M$, resulting in the following modification of the CT-NARX model

$$y_j^{n_i}(t) = f\left(y(t), \dots, y^{n_i-1}(t)\right) + k_j u_j(t) + e(t), \quad (4.17)$$

for $j = 1, \dots, M$ experimental signals corresponding to different constant input levels of 3OC₆HSL.

4.5.3 CT-NARX model with static input nonlinearity

The function $G(\cdot)$ was used to model the static nonlinear gain variation across input levels, which mapped the 3OC₆HSL input - $u(t)$ to the dynamic model input - $\tilde{u}(t)$, and the CT-NARX model was consequently modified to

$$y^{n_i}(t) = f\left(y(t), \dots, y^{n_i-1}(t)\right) + \tilde{u}(t) + e(t), \quad (4.18)$$

where $\tilde{u}(t) = G(u_*(t))$, $u_*(t) = \log_{10}(gu(t))$ (g is a scaling parameter discussed below). Due to the log spacing in the levels of 3OC₆HSL, the log transformation was applied to the scaled input $gu(t)$. The function $G(\cdot)$ was described by the basis function decomposition

$$\tilde{u}(t) = \sum_{j=1}^B w_j \psi_j(u_*(t)), \quad (4.19)$$

where $w_j \in \mathbb{R}$ is the j^{th} basis function parameter, B is the number of basis functions, and in this investigation the radial basis functions were used, specifically the squared exponential function,

$$\psi_j(u_*(t)) = \exp\left(-\frac{1}{2\sigma_j^2} \|u_*(t) - \varphi_j\|_2^2\right), \quad (4.20)$$

where φ_j and σ_j are the respective centres and widths of the j^{th} basis function. Basis functions were centred on the levels of the input data values $u_*(t)$ and the corresponding width parameters were heuristically tuned in the range $\sigma_j \in [1, 1.5] \forall j$.

4.5.4 Parameter estimation and structure detection of CT-NARX model

The basis function decomposition of the CT-NARX model shown in eqn(4.16) is advantageous because it is linear-in-the-parameters, hence least squares LS can be used for parameter estimation. The corresponding linear regression equation can be defined as

$$Y^{n_i} = \Phi\theta + \epsilon \quad (4.21)$$

where $Y^{n_i} = (\mathbf{y}_1^{n_i}, \dots, \mathbf{y}_M^{n_i})^\top$ is the model output vector of differential order n_i , ϵ is the model residual error vector, $\theta = (k_1, \dots, k_M, c_1, \dots, c_{N_\theta-1})^\top$, is the parameter vector, and $\Phi = [U \ Y]$ is the regression matrix where,

$$U = \begin{bmatrix} \mathbf{u}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{u}_M \end{bmatrix}, \quad (4.22)$$

$$Y = \begin{bmatrix} (\mathbf{y}_1^{(0)})^1 & (\mathbf{y}_1^{(1)})^1 & \dots & (\mathbf{y}_1^{(n_i-1)})^q \\ \vdots & \vdots & & \vdots \\ (\mathbf{y}_M^{(0)})^1 & (\mathbf{y}_M^{(1)})^1 & \dots & (\mathbf{y}_M^{(n_i-1)})^q \end{bmatrix}, \quad (4.23)$$

where $\mathbf{u}_j = (gu_j(t_0), \dots, gu_j(t_{N_y} - 1))^\top$, and $\mathbf{y}_j^{n_i} = (y_j^{n_i}(t_0), \dots, y_j^{n_i}(t_{N_y} - 1))^\top$ for $j = 1, \dots, M$. The LS estimate of the parameters is

$$\hat{\theta} = \Phi^+ Y^{n_i}, \quad (4.24)$$

where $\Phi^+ = (\Phi^\top \Phi)^{-1} \Phi^\top$. In order to improve the numerical conditioning of the regression matrix Φ , the input levels were rescaled using a gain g , where $g = 1 \times 10^{11}$.

The Y matrix contained a superset of model terms composed of polynomial transformations of $y(t)$ and its derivatives. The identification framework detected a parsimonious model structure composed of a reduced set of those model terms. In this investigation the number of model terms was relatively small (9 terms) and so the model structure was detected by an exhaustive search (see section 3.4.1) of all possible model term combinations ($2^9 = 512$).

There are a number of metrics that can be used to guide model selection MS, which are typically based on quantifying model accuracy, such as the mean sum

of squared error (MSSE) (shown in eqn(3.48)). The limitations of metrics such as the MSSE, which only incorporate model terms related to model accuracy, is that unnecessarily complex models can appear preferable. For example, if the number of model parameters corresponds to the number of data points, the model will precisely fit the observed data and will therefore appear preferable, although it is intuitively obvious that such a model will usually suffer from over-fitting and be unnecessarily complex. In order to compare models and obtain a parsimonious model, information criteria (IC) were used to obtain the optimal trade-off between model accuracy and model complexity - Akaike's and the Bayesian IC (Ljung, 1999)

$$AIC = \log\left(\left(1 + \frac{2N_\theta}{N_y}\right) \times \frac{1}{N_y} \sum_{t=0}^{N_y-1} e^2(t)\right) \text{ and} \quad (4.25)$$

$$BIC = \log\left(\left(1 + \frac{N_\theta \log(N_y)}{N_y}\right) \times \frac{1}{N_y} \sum_{t=0}^{N_y-1} e^2(t)\right) \quad (4.26)$$

respectively.

It has been suggested that the AIC does not always penalise model complexity sufficiently and hence the use of the BIC can lead to the selection of more compact models.

In implementation, the CT-NARX model was simulated using a first order Euler approximation. The basis function parameters for the static nonlinear input function $G(\cdot)$ were estimated using LS from the target data $\tilde{\mathbf{u}} = (k_1 u_1(t_0), \dots, k_M u_M(t_0))^T$, where $u_j(t_0)$, for $j = 1, \dots, M$, corresponded to the rescaled input levels of 3OC₆HSL.

4.5.5 Results and discussion

CT-NARX model of BBa_T9002 system

The number of possible candidate terms in a superset is normally very large, and so the model structure is detected using efficient search algorithms based on, for instance, the forward regression orthogonal FRO (Chen et al., 1989). The number of model terms in this investigation is very small (only 9 candidate model terms), therefore the model structure was detected by an exhaustive search of all possible models resulting from different model term combinations (a total of $2^9 = 512$). The AIC and BIC was used to detect the model with the optimal trade-off in terms

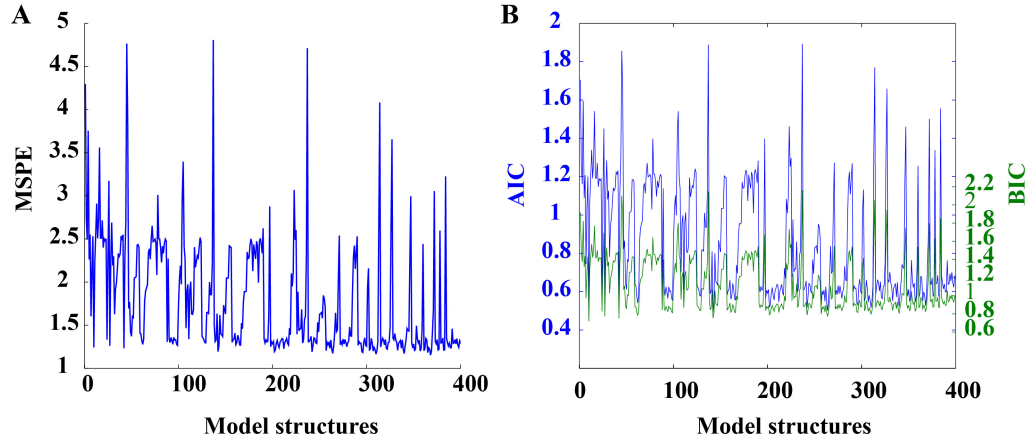


Figure 4.5: CT-NARX structure detection: (A) Mean squared prediction error (MSSE) for CT-NARX models with an MSSE < 5 and (B) Akaike and Bayesian information criteria (AIC and BIC respectively), optimal model with minimum AIC and BIC value is model structure 16. Note in both (A) and (B), the models are ordered by increasing complexity, i.e, number of model terms.

of maximum accuracy and minimal complexity. For MS, information criteria are preferable to the use of residual error metrics (*e.g.* mean sum of squared error, MSSE) for model comparison. The advantage is due to the ability of the ICs to penalise model complexity which MSSE cannot. The MSSE normally tends to decrease as the number of parameters increase. This is clearly illustrated in Figure 4.5A and B, where the MS result for experiment 1 is shown.

The "population-level" CT-NARX model of BBa_T9002 identified using the AIC and BIC was

$$\ddot{y}(t) = c_1 y^2(t) + c_2 y^3(t) + c_3 \dot{y}(t) + \tilde{u}(t),$$

where $y(t)$ was the model output signal rate of GFP expression, with associated parameters c_1 , c_2 and c_3 . The input term $\tilde{u}(t)$ was obtained from a static transformation $G(\cdot)$ of the input signal 3OC₆HSL, which was primarily used to describe the static switching effect in dynamics across linearly increasing levels of 3OC₆HSL (see above).

The error variance of the identified ERS model was 24.3%, whose prediction by simulation was shown to be inaccurate (Figure 4.3). However, in contrast, the CT-NARX model provided a much more accurate description of the BBa_T9002 system (an error variance of 0.2%) while retaining a simple model structure (com-

pare Figure 4.3 and Figure 4.6). The CT-NARX model was further validated and tested by predicting responses to additional input concentrations not used in the identification procedure. The predicted simulation results from using these intermediate inputs (3OC₆HSL input concentrations: $1 \times 10^{-9.5}$, $1 \times 10^{-8.5}$, $1 \times 10^{-7.5}$, $1 \times 10^{-6.5}$, $1 \times 10^{-5.5}$, $1 \times 10^{-4.5}$ M) demonstrated that the CT-NARX model behaved as expected (Figure 4.6B).

The main objective of this investigation is achieved, by identifying a "population-level" CT-NARX model that seeks to describe the same relationship as part of the ERS model: the input and output dynamic behaviour between 3OC₆HSL and GFP expression (eqn(4.7)). The model terms $y^2(t)$ and $y^3(t)$ are nonlinear model terms identified by the data-driven framework. These model terms have enabled the CT-NARX model to achieve better prediction power without significantly increasing the model complexity. However, the physical insight to these model terms biochemically is not yet known, but one for future consideration. In addition, the lack of interpretability of the model terms is not relevant to the utility of the model for use in design procedures, in which context it would appear that the identified CT-NARX model is highly preferable.

Consistency of identified model over a set of colonies

The data-driven framework was applied to each experimental dataset, 3 colonies with 3 replicates of each colony making a total of 9 experimental datasets: colonies 1 and 2 were used for identification of the CT-NARX model, and colony 3 was reserved for cross validation. A consistent model structure of the dynamic function was identified for all 6 experimental datasets across colonies 1 and 2. However, in 4 of the experimental datasets, the model structure of the CT-NARX model had little or no sensitivity towards the truncation point of the experimental data. While for the remaining 2 experimental data, there was some sensitivity detected, where slight differences occurred in the model structure depending on the exact time at which the experimental data was truncated. This occurrence is not abnormal, as sensitivity in data-driven modelling is common as time-domain descriptions are typically non-unique.

The parameters of the CT-NARX model were fairly identical across the experimental datasets - the variation observed was within an order of magnitude and consistent differences were noticed between colonies (Figure 4.7A-C). The CT-NARX model simulations were similar in accuracy to those shown in Figure 4.6, as illustrated by the similarity in MSSE (Figure 4.7D). The variations in all the parameters

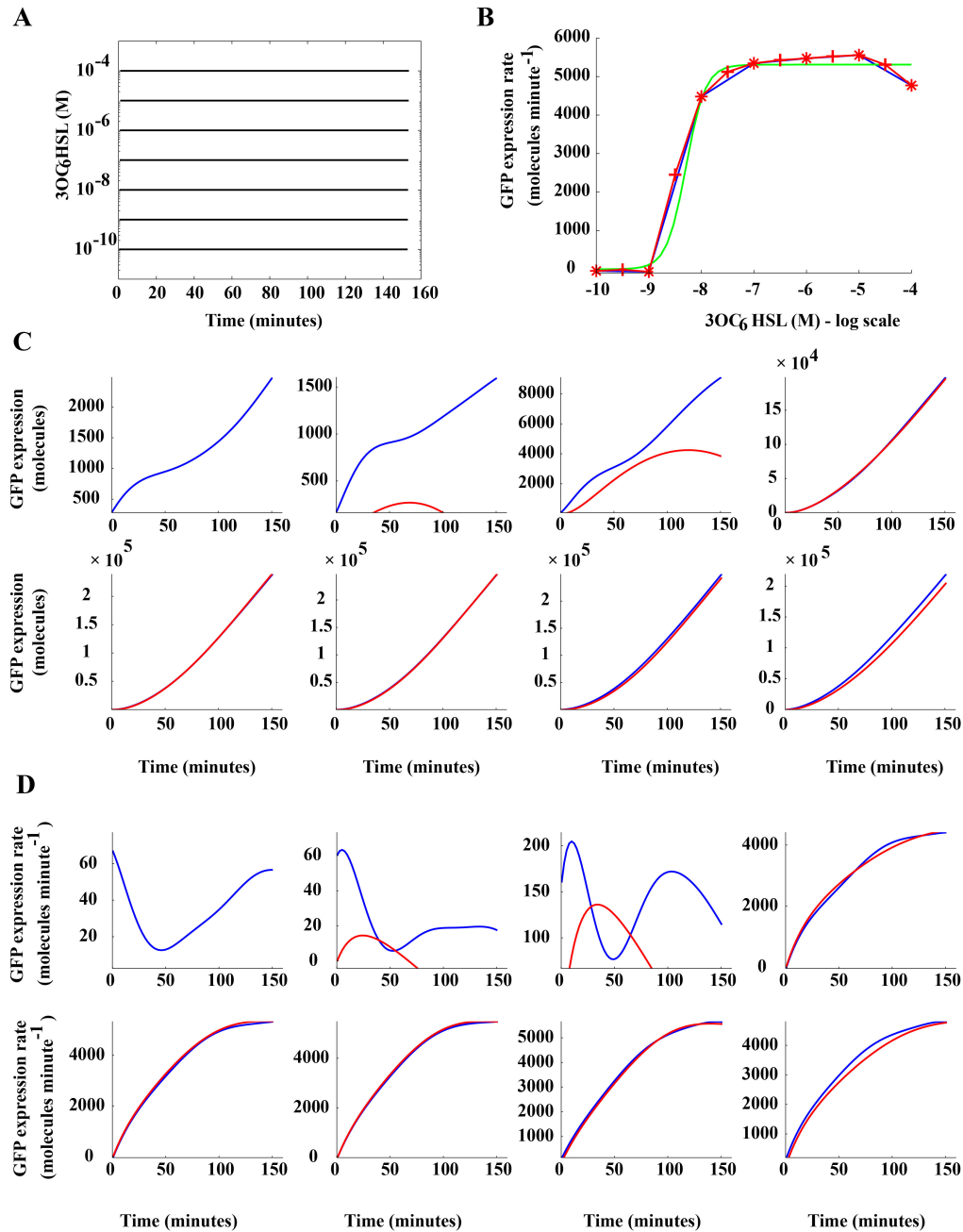


Figure 4.6: CT-NARX model simulation: (A) CT-NARX input signals, (B) Rate of change of GFP expression at the 150th minute (blue), the Hill equation prediction (green), the CT-NARX model prediction at observed input concentration (red stars), and the CT-NARX model prediction at interpolated input concentrations (red crosses), (C) Comparison of GFP expression (blue) and the CT-NARX prediction (red) and (D) Rate of change of GFP expression (blue) and CT-NARX model prediction (red). Note that the response corresponding to the lowest input level $3OC_6HSL = 0$ in (A) has been omitted because of the log transformation, and the plots in (C) and (D) corresponds to responses to increasing input level from left to right - top to bottom.

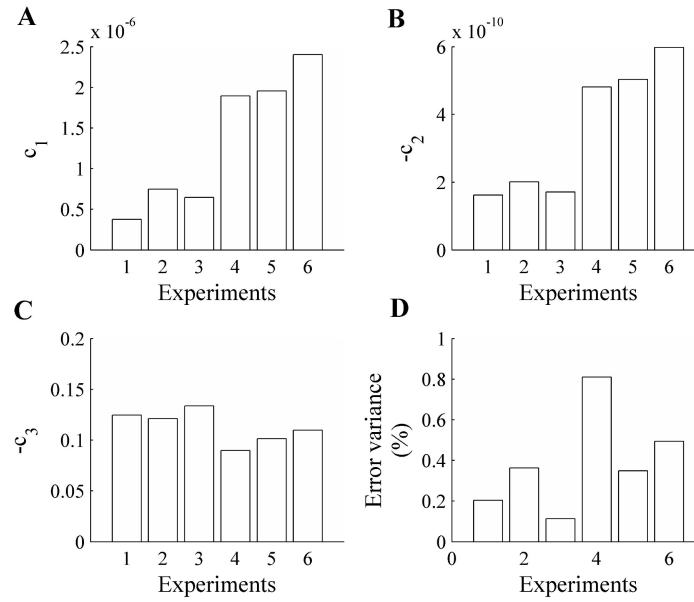


Figure 4.7: CT-NARX model identification across colonies and experimental data sets. (A-C) Estimates of CT-NARX model dynamic parameters c_1 , c_2 and c_3 . (D) CT-NARX model mean sum of squared error (MSSE) for each of 6 different experimental data sets where sets are grouped by colony: colony 1 comprises experiments 1-3; colony 2 comprises experiments 4-6.

Table 4.1: Mean and variability in CT-NARX model parameters across colonies

Parameters	c_1	c_2	c_3
Mean	1.34×10^{-6}	-3.53×10^{-10}	-0.1134
Standard deviation	8.46×10^{-7}	1.96×10^{-10}	0.016

of the CT-NARX model is summarised in Table 4.1.

Separate static input transformation functions $G_j(\cdot)$ was estimated, for $j = 1, \dots, 9$, corresponding to each experimental dataset. They were consistent, with some variability over scaling (Figure 4.8A) - an average static transformation was also estimated using all experimental datasets (Figure 4.8B). These variations observed in both static and dynamic parameters across experimental datasets are most likely due to population heterogeneity, which is an important characteristic to quantify. As discussed in Chapter 2, population heterogeneity arises due to variations at single-cell level - intrinsic and extrinsic noise.

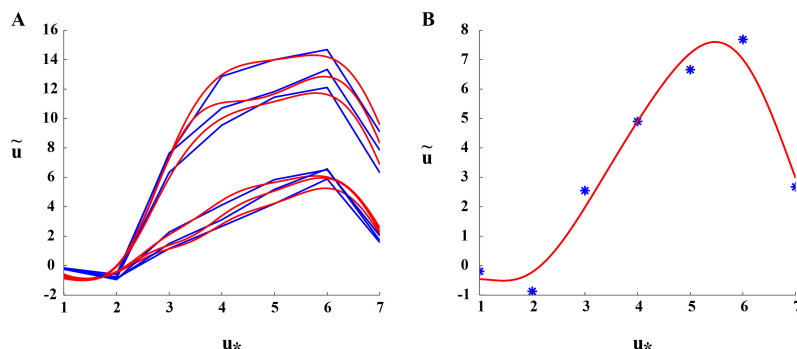


Figure 4.8: Static model of the input nonlinearity. The CT-NARX dynamic model input $\tilde{u}(t)$ was obtained from transforming $u_*(t) = \log_{10}(gu(t))$ through a static function $G(u_*(t))$, where $u(t)$ was the level of 3OC₆HSL. (A) Separate estimates of the static function $G(\cdot)$ (red) across 6 experimental data sets (blue) used for identification purpose and (B) Single estimate of the static function $G(\cdot)$ (red) using experimental datasets compared to the average of the experimental data curves in panel (A) (blue dots)

Model validation of a unified model

For design purposes, a single model description for the BBa_T9002 system across experimental datasets was provided, by using the mean values of the dynamics parameters in Table 4.1 along with the mean static transformation function shown in Figure 4.8B. The model prediction of the mean model for both training and validation experimental datasets, in describing the dynamics of the BBa_T9002 system was similar (Figure 4.9)

4.6 Summary and further discussion

In this chapter, a data-driven framework is proposed to identify a nonlinear black-box model for dynamic characterisation of genetic parts in biological systems. This framework has particular advantages for use in a top-down design in higher order systems. The identified CT-NARX model is compact, data-driven in both structure and parameters, and is part of a wider toolset of associated design and analysis methods. The framework was demonstrated on a transcriptional regulatory genetic part - BBa_T9002, for which an accurate dynamic model was obtained. The CT-NARX model was also benchmarked against dynamic and static biochemical models, which were based on an enzymatic reaction scheme. The enzymatic reaction scheme model was shown to be inaccurate and inconsistent with its associated simplified form - the Hill equation. In contrast to the reaction scheme model, the CT-NARX model provided an accurate dynamic description of the BBa_T9002 system whilst retaining a simple structure. On the basis of these results, the data-

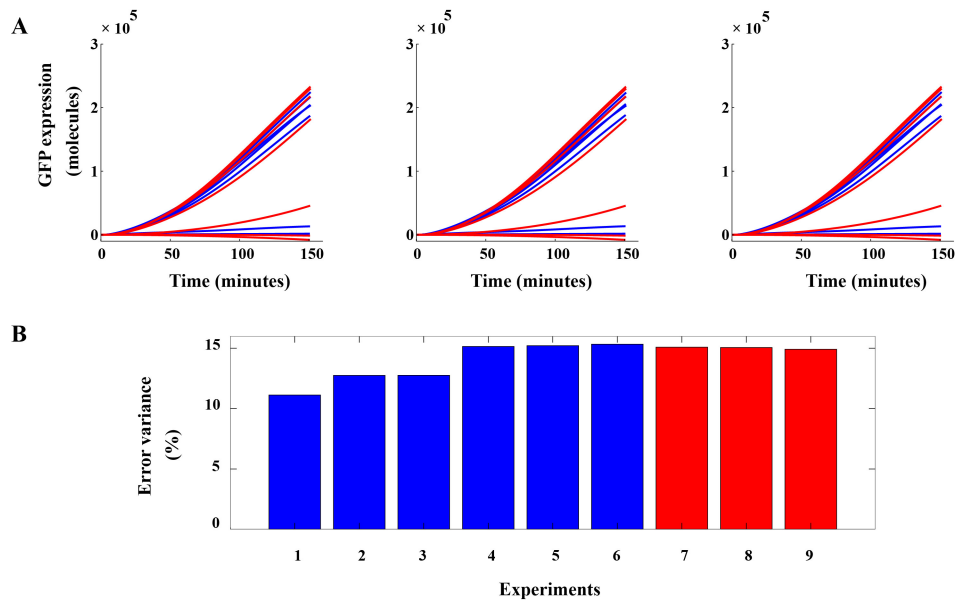


Figure 4.9: CT-NARX model prediction on validation data. (A) A single CT-NARX model with average parameter estimates was simulated (red) and compared to a reserved set of validation data (blue) and (B) The percentage prediction error variance from the averaged CT-NARX model using both estimation (blue) and validation (red) datasets.

driven framework offers great promise for use in the characterisation of synthetic genetic parts and further design procedures.

The observed dynamics of the BBa_T9002 system in this investigation were only collected up to the quasi steady state. The quasi steady state point in the cell growth curve is approximately the halfway point in the exponential growth stage (Figure 2.3). However dynamics beyond the quasi steady state point are also very important to characterise; in operational prototypes - designed chambers will contain cultures producing proteins at all stages of the growth curve, therefore dynamic models should be able to predict gene expression in all stages of the growth curve. The cell growth curves beyond the quasi steady state point are unlikely to be similar for different cultures (shown to be similar until quasi steady state in this investigation). Therefore gathering experimental data in which the dynamics of the system is observed through all stages of the cell growth is fundamentally important and the inclusion of the growth variable in the dynamic model becomes necessary. Observing the separate effects of both the cell growth and gene expression on the dynamics of the system will be explored in later chapters. Chapter 6 goes through the experimental setup, where the additional experimental data is collected.

The CT-NARX model implemented in this chapter, characterises the population heterogeneity through the noise term $e(t)$. However this does not fully and satisfactorily quantify and explain the variability phenomenon by different cell populations. Therefore the need to develop a better noise model or a better quantification methodology to capture the variability observed is of key interest in this thesis. In the following chapter (Chapter 5), a identification algorithm for CT systems is developed, which will be able to quantify the cell population variability that could be translated into features that would aid design procedures.

Chapter 5

A novel identification framework for continuous-time non-linear dynamic systems

5.1 Introduction

In the field of nonlinear system identification, there are many techniques available for obtaining discrete-time DT models (Baldacchino et al., 2012, Chen et al., 1989, Kukreja et al., 2004, Li et al., 2006, Piroddi and Spinelli, 2003). There are, however, far fewer techniques available for the identification of nonlinear continuous-time CT models. The prevalence of DT methods may be due to the ready availability of sampled data that can be directly used in nonlinear system identification algorithms as well as the typical desire to use the model in design for digital control systems, even when the process is inherently CT. However, there are a number of reasons why CT models are attractive for nonlinear system identification (see section 3.7): (i) they are easier to interpret and can to some extent facilitate physical understanding, (ii) they tend to be compact, (iii) they permit identification for irregularly sampled data and (iv) they exhibit more stability and less ill-conditioning (Garnier and Wang, 2008).

General approaches to CT nonlinear system identification include both direct and indirect methods: Either directly identifying the model in CT from sampled data, or indirectly by first identifying a DT model and then mapping it to the CT domain (Unbehauen and Rao, 1990). A major difficulty for direct CT system identification is estimating signal derivatives for use in regression-based parameter estimation methods (Garnier and Wang, 2008). The few techniques available for CT nonlin-

ear system identification overcome the signal derivative estimation problem in a variety of ways, *e.g.*: By use of delayed state-variable filters (Tsang and Billings, 1994), by Kalman smoothing (Coca and Billings, 1999) and by use of the delta-operator (Anderson and Kadiramanathan, 2007). Each of these approaches leads to an efficient regression-based estimation step and conveniently allows one to easily translate and apply DT model structure detection MSD methods to the CT nonlinear system identification task. The problem is that derivative estimation can be extremely difficult due to noise amplification and these methods suffer accordingly in noisy environments. In the earlier chapter (Chapter 4), the Kalman smoothing technique was applied to derive a CT nonlinear autoregressive model with exogenous input NARX. In extension, the need to appropriately quantify the observed variability in the experimental datasets could be partially achieved by deriving the CT nonlinear autoregressive moving average model with exogenous input NARMAX; inclusive of a more detailed noise model. However, its identification will also be hampered by the noisy derivative estimation. Therefore, in this chapter an alternative identification approach is pursued.

The aim of this chapter is the development of a novel algorithm for direct CT nonlinear system identification that is more robust to noisy signals. The algorithm proposed is focused on a simulation approach, making use of the output-error model structure as opposed to the equation-error structure. The key advantage of a simulation approach is that parameter estimation and MSD do not require estimation of signal derivatives. This is because there is no regression-based estimation step within the simulation framework. Piroddi and Spinelli (2003) proposed a simulation-based identification algorithm for DT nonlinear systems but in that case parameter estimation was performed by linear regression, which for CT system identification would again incur the disadvantage of signal derivative estimation. Piroddi and Spinelli (2003) reported advantages for the simulation term selection under some restrictive conditions such as non-persistently exciting signals and fast sampling (Billings, 2013), that should carry over to the CT case. However, an entirely new algorithm is required that performs the step of parameter estimation using the simulated signals.

The choice of estimation criterion for nonlinear system identification ranges across least squares LS, maximum likelihood and Bayesian approaches (Ljung, 1999, Peterka, 1981). In both control engineering and in other areas such as the life sciences where nonlinear system identification is now more commonly applied (Anderson et al., 2010, Krishnanathan et al., 2012, Kukreja et al., 2003), systems can have wide

variation in dynamic behaviour (Chapter 4 is also an example). Hence, obtaining a parameter distribution is typically of more interest than a single best-fit estimate. The choice of estimation criterion is therefore guided here by the requirement of characterising uncertainty, resulting in a Bayesian approach.

Within the area of Bayesian estimation, computational methods are gaining popularity due to advances in computational processing power (Baldacchino et al., 2013, Ninness and Henriksen, 2010). A particular approach, known as ABC (Beaumont, 2010, Tavaré et al., 1997, Toni et al., 2009) is a rejection sampling algorithm well suited to the nonlinear estimation problem encountered in CT system identification. This is because ABC is a likelihood-free Bayesian estimation method, which bypasses the need to analytically derive the likelihood function - a complicated task for nonlinear CT models. Instead, the likelihood is numerically approximated via model simulations that are deemed close, in some sense, to the observed data. In practice, model simulations are performed with randomly drawn parameter sets, and these sets are rejected if they lead to simulations that are outside of some distance threshold.

Given the parameter estimation approach of ABC, the question arises of how to develop a MSD step for the nonlinear CT model. Firstly, the ABC algorithm naturally builds a numerical representation of the parameter distributions. Conveniently, Kukreja et al. (2004) have developed a NARMAX model term selection method based on exploiting estimates of parameter distributions (obtained from bootstrapping). Hence, the inspiration of this chapter's approach to term selection comes from Kukreja et al. (2004), where terms are pruned from a model superset using a significance test: parameter ranges are checked to see if they include zero - corresponding terms are pruned from the model. This algorithm is extended here to account for less sensitive terms that fail the basic significance test. The main result is a new algorithm for CT nonlinear system identification, in a computational Bayesian framework with a simulation focus.

In this chapter, a system identification framework for CT nonlinear systems is developed, for the first time using a computational Bayesian approach. The framework has two main advantages over existing comparable techniques: (i) parameter distributions are naturally generated, giving the user a clear description of uncertainty and (ii) the method is well suited to noisy signals because it avoids the need to estimate signal derivatives. The ABC algorithm is used to estimate model parameters in a simulation-based framework. Term selection is performed by

judging parameter significance using parameter distributions that are naturally generated as part of the ABC procedure. The results from numerical examples demonstrate that the method performs well in noisy scenarios, especially in comparison to competing techniques that rely on signal derivative estimation.

5.2 Parameter estimation by approximate Bayesian computation

There are several systems which are very complex and also exhibit high nonlinearity such as biological systems. The need to derive models that could describe these systems are useful. It is also useful to estimate the parameters of these models in order to simulate them, and hence obtain simulated predictions of the system dynamics under consideration. The simulated predictions could be cross-validated against observed outputs, allowing the performance assessments of these models and their parameters respectively. The assessments are preferred to be non-subjective and clear.

In the Bayesian framework, the inference procedure tends to estimate the probability of a given set of parameters in order to obtain a observed output, thereby computing the likelihood function. The computation of the likelihood function in most cases are very difficult. Typically, Bayesian inference involves the estimation of a conditional distribution which computes the posterior distribution,

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{y})}, \quad (5.1)$$

where $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function, $\pi(\boldsymbol{\theta})$ is the prior distribution and $p(\mathbf{y})$ is the marginal likelihood. However, this is difficult to compute because the marginal likelihood

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \quad (5.2)$$

is mostly a high dimensional integral (Beaumont, 2010). To the contrary, coding a simulation program for a model of the process described by $p(\mathbf{y}|\boldsymbol{\theta})$ is much more straightforward and easy. The ABC estimation method exploits this situation by sampling directly from the posterior distribution to directly obtain an estimate of the conditional distribution $p(\boldsymbol{\theta}|\mathbf{y})$ (Beaumont et al., 2002, Tavare et al., 1997). This allows for a principled way of comparing predicted data and experimental data, and most importantly avoids the computation of the likelihood function but still carries all the advantages of the Bayesian framework. In the simplest

form of ABC, given a model for which parameter values are simulated from a prior distribution, which is used in simulating the model, and hence the predicted output is compared with the experimental data using a distance metric. If the value of the distance metric is below a chosen threshold, the parameter set is picked as a sample from the posterior distribution.

5.2.1 Rejection sampling

The basic ABC algorithm is termed the rejection sampling (the modified version presented in (Rubin, 1984)). The rejection sampling algorithm is:

1. Draw $\theta^* \sim \pi(\theta)$
2. Simulate $\mathbf{y}^* \sim p(\mathbf{y}|\theta^*)$
3. Reject θ^* if $d(S(\mathbf{y}^*), S(\mathbf{y})) > \epsilon$

where $S(\cdot)$ describes summary statistics (that may be vector-valued), $d(\cdot)$ is a distance measure between simulated and observed values, and ϵ is a threshold value. There are several ways to visualise the posterior distribution but can be simply done using a histogram. A smoothed version of the distribution can be achieved using kernel density methods, that allows the computation of the distribution's mode, mean or quantiles. The summary statistic is used to reduce the dimensionality of the observed data, when it involves many uni- or multi- variate measurements. This process usually tends to lose information.

5.2.2 Inefficiency of the basic ABC

The basic ABC is very inefficient in practice. It is always desirable that the threshold value, ϵ , is relatively small so that the parameter samples accepted are correspondingly good as far as possible. Moreover, if a small value is assigned to ϵ , a large number of sampled parameters will be rejected. This gives rise to an inefficient solution, as there is always a computational limitation of the number of parameters that could be sampled from the prior realistically. This is even more so, if the prior distribution is very different from the posterior distribution.

There are modified approaches to the basic ABC algorithm which counter the inefficiency problem. The regression-based conditional density estimation approach was introduced in Beaumont et al. (2002). It introduces a weight for each accepted parameter sample, in which the sample that corresponds to the lowest error threshold is weighted higher in comparison to the others. It also applies a

regression-adjustment step, where each accepted parameter is described by a linear regression equation that is used to obtain an adjustment, thereby estimating the marginal posterior distribution of the parameter. Theoretically, the weighting and adjustment steps increase the efficiency, however, the approach only samples from the prior distribution therefore the quality of the results highly rely on informative prior. A Markov chain Monte Carlo approach of the ABC was introduced in Marjoram et al. (2003). It produces a growing Markov chains of the parameter samples until a stationary distribution is attained to describe the true posterior distribution of the parameters. The drawback of this approach is the correlation of the accepted parameter samples and the possibility of chains being stuck in low density regions resulting to slow convergence. An alternating and efficient approach is the ABC-sequential Monte Carlo (SMC) approach. This is discussed later on and the justification of its efficiency is given.

5.3 Model definition and parameter estimation

5.3.1 Continuous-time nonlinear model representation

The model structure considered in this chapter is CT nonlinear output error NOE model. The observed output $y(t) \in \mathbb{R}$ of a CT-NOE process can be represented as

$$z^{n_i}(t) = f(\mathbf{S}(t)), \quad (5.3)$$

$$y(t) = z(t) + e(t), \quad (5.4)$$

where $z(t) \in \mathbb{R}$ is the unknown noise-free system output, $u(t) \in \mathbb{R}$ is the known system input, $z^{n_i}(t) \in \mathbb{R}$ indicates the n_i^{th} derivative of $z(t)$, and the measurement noise $e(t)$ is assumed to be zero-mean white noise. The function $f(\cdot)$ describes the dynamics of the nonlinear CT process and $\mathbf{S}(t) \in \mathbb{R}^{2n_i}$ is the vector of input and output derivatives,

$$\mathbf{S}(t) = \left(z(t), \dots, z^{n_i-1}(t), u(t), \dots, u^{n_i-1}(t) \right). \quad (5.5)$$

The nonlinear function $f(\mathbf{S}(t))$ can be decomposed and represented by a linear sum of basis functions $\phi_j(\mathbf{S}(t))$ which can have varying forms including wavelet, polynomial or radial functions,

$$f(\mathbf{S}(t)) = \sum_{j=1}^{N_\theta} \theta_j \phi_j(\mathbf{S}(t)), \quad (5.6)$$

where N_θ is the number of model terms and $\theta_j \in \mathbb{R}$ is the parameter associated with basis function $\phi_j(\cdot)$.

5.3.2 Parameter estimation by ABC-SMC

A shortcoming of the basic ABC algorithm is the low acceptance rate when the prior distribution is very different to the true posterior (as discussed above). A low acceptance rate would require many simulations to adequately represent the posterior distribution. To increase the computational efficiency of ABC, therefore, more sophisticated approaches have been developed (see above for more details) (Beaumont, 2010). One such method is the ABC-SMC algorithm (Sisson et al., 2007, 2009, Toni et al., 2009), which has proved effective in dynamic systems modelling (Holmes et al., 2012, Liepe et al., 2012).

The main idea of the ABC-SMC algorithm is to iterate population estimates generated by ABC, gradually decreasing the error tolerance ϵ_l at each iteration l . The posterior distribution at iteration l becomes the sampled prior distribution at $l + 1$. Hence, the ABC-SMC algorithm reaches the target posterior in a sequential manner.

The error threshold sequence is chosen so that it decreases at each iteration, hence $\epsilon_1 > \dots > \epsilon_L$, where L is the number of iterations. The first and final thresholds can be tuned by performing the basic ABC estimation algorithm for N_s samples and setting $\epsilon_1 = 2d_{min}$ and $\epsilon_L = 1.2d_{min}$, where d_{min} denotes the minimum of the vector of all N_s distance measures. The ABC-SMC algorithm is described for the nonlinear CT model parameter estimation in Algorithm 5.1. Further details of the implementation is given in the following subsection.

5.3.3 Implementing ABC-SMC for nonlinear continuous-time model

For the nonlinear CT model, the prior distribution of the parameters was defined here as a uniform distribution. In the absence of specific information on the prior, it was scaled using the LS parameter estimate obtained from the method of Coca and Billings (1999). Hence, the prior distribution was defined as

$$\pi(\boldsymbol{\theta}) \sim U(-2\gamma, 2\gamma), \quad (5.7)$$

with range parameter γ ,

$$\gamma = (\Psi^\top \Psi)^{-1} \Psi^\top \hat{\mathbf{y}} \quad (5.8)$$

Algorithm 5.1 Parameter Estimation by ABC-SMC

Require: number of iterations L , number of parameter samples N_s ,

prior $\pi(\boldsymbol{\theta})$ and error sequence $\epsilon_1 > \dots > \epsilon_L$

for $l = 1$

for $j = 1 : N_s$

draw $\boldsymbol{\theta}_j^* \sim \pi(\boldsymbol{\theta})$ and simulate $\mathbf{y}_j^* \sim p(\mathbf{y}|\boldsymbol{\theta}_j^*)$

until $d(S(\mathbf{y}_j^*), S(\mathbf{y})) \leq \epsilon_1$

end for

set each weight $w_j^1 = \frac{1}{N_s}$

end for

for $l = 2 : L$

for $j = 1 : N_s$

sample $\boldsymbol{\theta}_j^*$ from $\boldsymbol{\theta}^{l-1}$ with probabilities w^{l-1}

perturb $\boldsymbol{\theta}_j^*$ to obtain $\boldsymbol{\theta}_j^{**} \sim \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$

simulate $\mathbf{y}_j^* \sim p(\mathbf{y}|\boldsymbol{\theta}_j^{**})$ until $d(S(\mathbf{y}_j^*), S(\mathbf{y})) \leq \epsilon_l$

end for

set each $\boldsymbol{\theta}_j^l = \boldsymbol{\theta}_j^{**}$

set each $w_j^l = \frac{\pi(\boldsymbol{\theta}_j^l)}{\sum_{i=1}^{N_s} w_i^{l-1} \mathcal{L}(\boldsymbol{\theta}_i^{l-1}|\boldsymbol{\theta}_j^l)}$, and normalise

end for

Note: parameter samples are denoted as $\boldsymbol{\theta}^*$, and $\boldsymbol{\theta}^{**}$ after perturbation. \mathcal{L} is a parameter perturbation kernel (uniform random walk).

where

$$\Psi = \left(\boldsymbol{\psi}(t_0)^\top, \dots, \boldsymbol{\psi}(t_{N_y-1})^\top \right)^\top \quad (5.9)$$

$$\boldsymbol{\psi}(t) = \left(\hat{y}(t), \dots, \hat{y}^{n_i-1}(t), u(t), \dots, u^{n_i-1}(t) \right) \quad (5.10)$$

$$\hat{\mathbf{y}} = \left(\hat{y}^{n_i}(t_0), \dots, \hat{y}^{n_i}(t_{N_y-1}) \right)^\top \quad (5.11)$$

where the derivative estimates of $y(t)$ were obtained from a Kalman smoothing algorithm, described in Coca and Billings (1999). The model simulation step was performed by deterministic simulation of the model defined in eqn(5.3 and 5.4), using a fourth order Runge-Kutta method. The distance measure of simulations from observations was obtained from the sum-of-squared errors,

$$d = \sum_{j=0}^{N_y-1} (y(t_j) - y^*(t_j))^2. \quad (5.12)$$

The L_2 norm used here for $d(\cdot)$ is suited to normally distributed noise but for other types of noise it would be possible to use an alternative, for example an L_1 norm for Laplacian noise or an L_∞ norm for uniform noise.

5.4 Nonlinear continuous-time model identification framework

In this section an identification framework is developed for the nonlinear CT model. First, a simple one-stage approach to MSD is derived, based on a parameter significance test. The significance test makes use of the parameter distributions naturally generated as a byproduct of the ABC-SMC algorithm. This works effectively for terms with sensitive parameters (explained in more details below). For terms with less sensitive parameters, a two-stage algorithm is derived that follows the significance test with an information criterion test.

5.4.1 One-stage model structure detection

The ABC-SMC algorithm naturally generates parameter distributions as part of the estimation procedure. Here, this feature is exploited by developing a MSD algorithm that makes direct use of these distributions. Similarly to the approach of Kukreja et al. (2004), a significance test is used to prune ‘false’ parameters from a superset of model terms, where significance is determined from the parameter

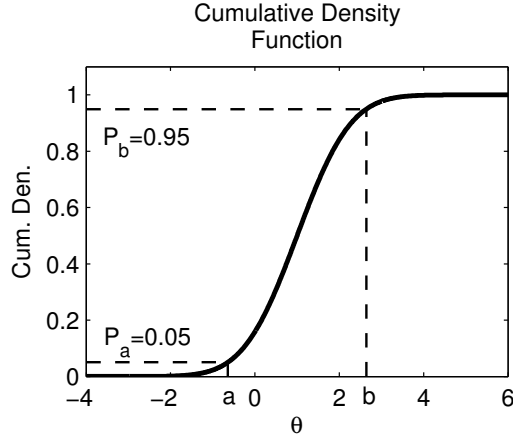


Figure 5.1: Term selection via the cumulative density function. The cumulative density function for a parameter θ is constructed by the ABC-SMC estimation algorithm. The model term is rejected if zero lies between the limits corresponding to the 5% and 95% probability levels, *i.e.* $a \leq 0 \leq b$.

distributions using a quantile test.

The quantile test selects parameter estimates that cannot be distinguished from zero: The test finds the intervals from the cumulative distribution of the parameters. First, the 5% and 95% intervals are defined

$$P_a = Pr(\theta \leq a) = 0.05, \quad P_b = Pr(\theta \leq b) = 0.95, \quad (5.13)$$

where a and b are constants that correspond to the probability levels 0.05 and 0.95 respectively. Then the quantile test is performed by checking if zero lies between the 5% and 95% intervals, *i.e.* $a \leq 0 \leq b$ (see Figure 5.1). The quantile test is used here unlike the percentile test used by Kukreja et al. (2004) because the posterior distributions obtained in the ABC-SMC framework can be skewed.

The algorithm proceeds as follows. An initial superset of model terms is defined, \mathcal{M}_0 , of cardinality $N_0 = |\mathcal{M}_0|$, where model terms correspond to basis functions ϕ_j in eqn(5.6). All parameters of the model terms in the set \mathcal{M}_0 is estimated using Algorithm 5.1 (ABC-SMC). The terms of model \mathcal{M}_1 is selected by forming quantile intervals a_j and b_j for each parameter θ_j and pruning the term if zero lies in the interval. The one-stage MSD algorithm is fully described in Algorithm 5.2.

Algorithm 5.2 One-stage model structure detection

Require: derivative order n_i and polynomial order q

Define: superset of model terms \mathcal{M}_0 , where $N_0 = |\mathcal{M}_0|$

Run: **Algorithm 1** for \mathcal{M}_0 (estimate parameters)

Initialise: $\mathcal{M}_1 = \mathcal{M}_0$

for $j = 1 : N_0$ (quantile test)

if $a_j \leq 0 \leq b_j$

 discard model term ϕ_j from \mathcal{M}_1

else

 retain model term ϕ_j in \mathcal{M}_1

end if

end for

Run: **Algorithm 1** for \mathcal{M}_1 (re-estimate parameters)

5.4.2 Two-stage model structure detection

In this section an enhanced two-stage MSD algorithm is described. To motivate this enhancement, firstly, it should be noted that an advantage of the one-stage quantile test is its computational efficiency. The quantile test is efficient because it only requires one pass through the ABC-SMC algorithm. However, a disadvantage of the quantile test is that it does not directly assess the performance of the model simulations. Therefore, in this section a second stage to the algorithm for term selection is developed that directly measures simulation performance using the Bayesian information criterion (BIC).

The two-stage algorithm proceeds as follows. At stage one the set of model terms \mathcal{M}_1 using the quantile test is obtained: The quantile test is only used on terms that pass an initial sensitivity test, where sensitivity is assessed as a parameter variance less than some constant β (here $\beta = 1$) - defined as pool 1 terms, the set \mathcal{P}_1 . Terms with parameter variance greater than β are defined as pool 2 terms, the set \mathcal{P}_2 . Terms in pool 2 have an ambiguous contribution to the model and therefore require further testing in the second stage. The ambiguity arises because model terms whose parameters are less sensitive have slow convergence towards their true parameter values, these model terms could be either correct or spurious terms. These model terms usually have poor prior initialisation, which tend to have wide uniform distributions. The wide uniform distribution occurs because: (i) the initialisation of the prior is done through the LS estimate using (Coca and Billings, 1999) approach and due to the large number of possible model terms, the estimation is an ill-conditioned problem, therefore some parameter estimates (for both correct and spurious model terms) could be misleadingly large in comparison

Algorithm 5.3 Two-stage model structure detection

Require: derivative order n_i and polynomial order q

Define: superset of model terms \mathcal{M}_0 , where $N_0 = |\mathcal{M}_0|$

Run: **Algorithm 1** for \mathcal{M}_0

Initialise: $\mathcal{P}_1 = \emptyset$ and $\mathcal{P}_2 = \emptyset$

for $j = 1 : N_0$ (determine \mathcal{P}_1 and \mathcal{P}_2)

if $\text{variance}(\theta_j) \leq \beta$

 allocate term ϕ_j to \mathcal{P}_1

else

 allocate term ϕ_j to \mathcal{P}_2

end if

end for

Set: $N_1 = |\mathcal{P}_1|$ and $N_2 = |\mathcal{P}_2|$

Initialise: $\mathcal{M}_1 = \mathcal{P}_1$

for $j = 1 : N_1$ (quantile test)

if $a_j \leq 0 \leq b_j$

 discard model term ϕ_j from \mathcal{M}_1

else

 retain model term ϕ_j in \mathcal{M}_1

end if

end for

Run: **Algorithm 1** for \mathcal{M}_1 (re-estimate parameters)

Order terms in \mathcal{P}_2 by descending Cha-Srihari metric

Initialise: $\mathcal{M}^{(0)} = \mathcal{M}_1$

for $j = 1 : N_2$ (BIC test for ordered \mathcal{P}_2)

 Form model $\mathcal{M}^{(j)}$ by adding term $\mathcal{P}_2(j)$ to $\mathcal{M}^{(j-1)}$

Run: **Algorithm 1** for $\mathcal{M}^{(j)}$ (re-estimate params)

if $\text{BIC}(\mathcal{M}^{(j)}) < \text{BIC}(\mathcal{M}^{(j-1)})$

 retain $\mathcal{P}_2(j)$ in $\mathcal{M}^{(j)}$

else

 break

end if

end for

Set: $\mathcal{M}_2 = \mathcal{M}^{(j)}$

to their true value and (ii) the true estimate of the parameters are actually large. However, the common feature of model terms and their respective parameters (either categorised under (i) or (ii)) are their low contribution (less sensitive) towards the dynamics of the system, hence most samples obtained from their wide prior are uninformative. Therefore in the second stage, the final set of selected model terms \mathcal{M}_2 is obtained by iteratively testing pool 2 terms using the BIC, after first ordering pool 2 using the Cha-Srihari metric (defined below) (Cha and Srihari, 2002).

The key step in the two-stage algorithm is the ordering of unselected terms by use of the Cha-Srihari distance metric (Cha and Srihari, 2002). The purpose of using Cha-Srihari is to detect which model parameter distributions have evolved the most from their uniform prior. It is assumed that the estimated parameter distributions that least resemble their uniform prior contribute the most to describing system dynamics. The ordering of model terms makes the search through pool 2 much more efficient than taking the unselected terms at random (in a way avoiding ill-conditioning). The Cha-Srihari distance, $D(A, B)$, measures how much effort it takes to transform a reference histogram, A (the prior), to a target histogram B (the posterior),

$$D(A, B) = \sum_{i=1}^{N_h} |s_i|, \text{ for } i = 1, \dots, N_h, \quad (5.14)$$

where $s_i = \sum_{j=1}^i r_j$, for $i = 1, \dots, N_h$; $r_i = A_i - B_i$, A_i and B_i are bar sizes of histograms A and B respectively, and N_h is the number of bars. Here $N_h = 4$ is set with bar centers in between the intervals $[-2\gamma, -\gamma, 0, \gamma, 2\gamma]$. The set of pool 2 terms is sorted in descending order of Cha-Srihari measure, which are then searched in order using the BIC. The two-stage MSD algorithm is fully described in Algorithm 5.3.

5.4.3 Derivative order model selection

Identifying the correct derivative order, n_i , of the nonlinear CT model is an important issue to address. Here the Bayes factor criterion is used to identify the correct derivative order, as this naturally fits with the ABC framework. ABC can be used in model selection by allocating competing models an index, and then treating this index selection as a parameter estimation problem (Toni et al., 2009). The Bayes factor for comparing evidence supporting two models with different derivative order \mathcal{M}_i and \mathcal{M}_j is

$$B_f(i, j) = \frac{p(\mathcal{M}_i|\mathbf{y})/p(\mathcal{M}_j|\mathbf{y})}{p(\mathcal{M}_i)/p(\mathcal{M}_j)}. \quad (5.15)$$

which for uniform priors simplifies to $B_f(i, j) = \frac{p(\mathcal{M}_i|\mathbf{y})}{p(\mathcal{M}_j|\mathbf{y})}$. In practice, for derivative order selection, Algorithm 5.3 is run independently for models of different derivative order and then models are compared using the Bayes factor in a final run of the basic ABC algorithm.

5.5 Multi-core processing for fast ABC

The most time consuming steps in the ABC algorithm are the model simulations, which are typically performed many thousands of times. The model simulations are inherently parallelisable, due to their independence. Therefore, the ready availability of multi-core desktop machines were exploited to decrease the computation time. Specifically, custom algorithms in MATLAB was used to implement the parallelised ABC-based algorithms in conjunction with the Parallel Processing Toolbox (the *parfor* function, which automatically divides a task across available processors). An advantage of this approach is the ease of implementation, which does not require specialist knowledge of parallel programming. The decrease in computation time for simulating 1000 simulations is illustrated in Figure 5.2, which shows a 10 fold decrease in computation time using 12 cores in comparison to a core. Similar performance enhancements could be obtained with graphics processing units (GPUs) (Henriksen et al., 2012, Lee et al., 2010), which would require more specialist implementations.

5.6 Results

This section is divided into the following subsections:

1. The parameter estimation of the Van der Pol oscillator (VDPO) system corrupted by measurement noise - zero-mean Gaussian random noise with SNR of 20dB and 10dB.
2. The demonstration of the one-stage model structure detection algorithm on the VDPO system corrupted by measurement noise - zero-mean Gaussian random noise with SNR of 10dB. The result is compared and benchmarked against the derivative continuous-time method dCTM approach.
3. The demonstration of the two-stage model structure detection and derivative order model selection algorithms on a test system with identifiability

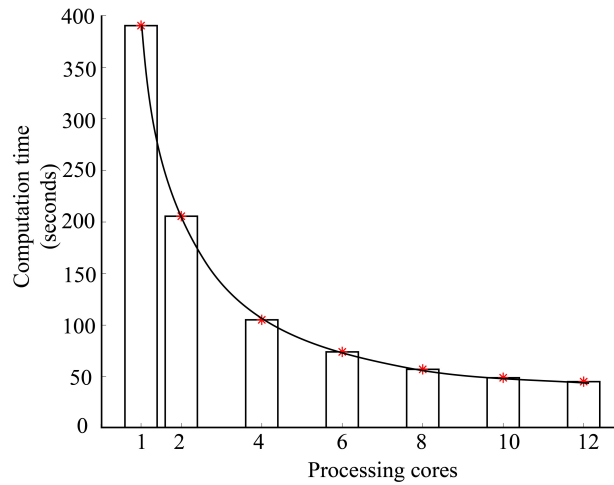


Figure 5.2: Computation time of 4th order Runge-Kutta simulation algorithm. 1000 random VDPO simulations were repetitively simulated using the *parfor* function with varying processing cores.

problem, which is corrupted by measurement noise - zero-mean Gaussian random noise with SNR of 20dB and 10dB. The result is also compared and benchmarked against the dCTM approach.

5.6.1 Case 1: parameter estimation of VDPO system

To investigate the performance of the ABC-SMC algorithm for parameter estimation, it was applied to the well known Van der Pol oscillator (VDPO) system, with increasing measurement noise. The VDPO system used was

$$\ddot{z}(t) = \theta_1 z(t) + \theta_2 \dot{z}(t) + \theta_3 u(t) + \theta_4 z^2(t) \dot{z}(t), \quad (5.16)$$

$$y(t) = z(t) + e(t), \quad (5.17)$$

where $e(t)$ was defined as zero-mean Gaussian random noise with variance (i) $\lambda^2 = 0.04$ for SNR of 20dB and (ii) $\lambda^2 = 0.25$ for SNR of 10dB. The parameter vector was set to $\theta = (-1, 0.2, 1, -0.2)$. The excitation signal was set to a zero-mean uniform random sequence in the range $(-20, 20)$ band-limited to 20 Hz. In implementing the parameter estimation the parameter size was set to $N_s = 200$ and the number of population iterations to $L = 10$.

If the assumption is that the posterior distributions are unimodal and symmetric, estimates from a distribution can be made on the basis of the mean in a minimum

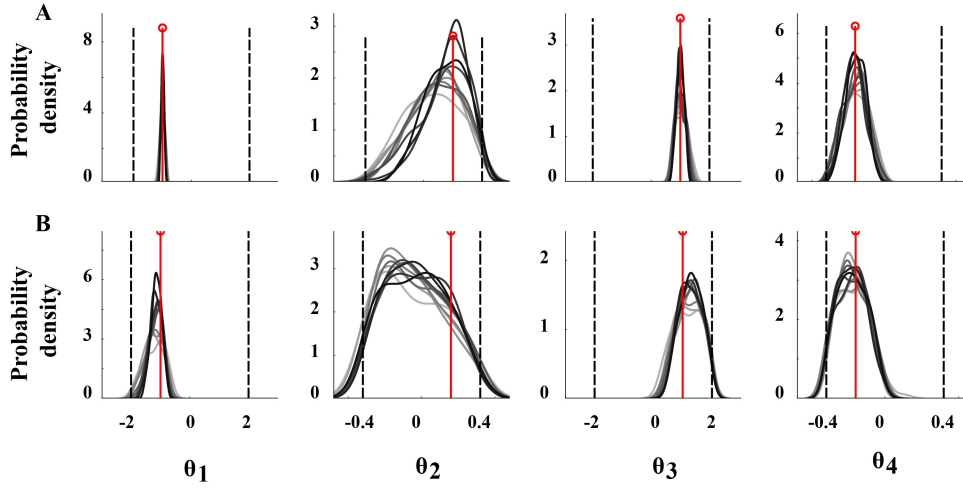


Figure 5.3: Parameter estimation of VDPO system using ABC-SMC (Algorithm 5.1). True parameters are shown as red stem plots and the black dotted lines indicates the prior. Estimated sample distributions are shown over 10 iterations of the ABC-SMC procedure on VDPO system with measurement noise for SNR of 20dB - (A) and 10dB - (B) measurement noise. Iteration 1 in grey while iteration 10 in black.

Table 5.1: Estimated parameters of the VDPO system.

	SNR = 20 dB	SNR = 10 dB
True System	ABC-SMC	ABC-SMC
$-1.00z(t)$	$-1.10z(t)$	$-1.56z(t)$
$0.2\dot{z}(t)$	$-0.21\dot{z}(t)$	$-0.39\dot{z}(t)$
$1.00u(t)$	$0.72u(t)$	$0.48u(t)$
$-0.2z^2(t)\dot{z}(t)$	$-0.36z^2(t)\dot{z}(t)$	$-0.39z^2(t)\dot{z}(t)$

Table 5.2: Identified models of the VDPO system. ABC1 refers to the one-stage ABC identification method.

True System	dCTM	ABC1
$-1.00z(t)$	$-1.06z(t)$	$-0.94z(t)$
$0.2\dot{z}(t)$	–	$0.19\dot{z}(t)$
$1.00u(t)$	$0.49u(t)$	$1.13u(t)$
$-0.2z^2(t)\dot{z}(t)$	$-0.01z^2(t)\dot{z}(t)$	$-0.26z^2(t)\dot{z}(t)$
–	$-0.01u(t)\dot{z}(t)$	–

mean squared error sense or mode in a maximum a posterior probability sense. The mode is often considered a more appropriate measure, particularly when it is used in a Bayesian framework. Mean estimate on the other hand is easier to compute than the mode. If the distribution is unimodal and symmetric then the mode and the mean will coincide.

One could observe that the parameter’s posterior distributions of the VDPO system whose measurement noise for SNR of 20dB, are much narrower in comparison to distributions obtained when a SNR of 10dB was used (Figure 5.3). However, the parameters of the VDPO system was accurately estimated for both measurement noise level (Table 5.1), indicating the robustness of the ABC-SMC procedure defined in Algorithm 5.1.

5.6.2 Case 2: model structure detection of VDPO system

To investigate the performance of the one-stage model structure detection algorithm, the VDPO system was used (eqn(5.16 and 5.17)) with the same parameter vector mentioned in the above subsection. However, the results shown below only entails that of the VDPO system corrupted by measurement noise with SNR of 10dB. The nonlinear order was set to $q = 3$ and the derivative order to $n_i = 2$. The results were compared to the dCTM approach that uses Kalman smoothing to estimate signal derivatives developed for CT systems by Coca and Billings (1999). Note, for the dCTM approach the tuning parameters for smoothing were chosen as discussed in section 3.7.2.

The contribution of each model term in the VDPO system towards the output dynamics is fairly high, therefore they do not tend to have identifiability issues. This could be illustrated by computing the unexplained variance (relative to the regressed output) of each term: $\sim (124.18, 126.80, 2.64, 127.01)$ that corresponds to $z(t), \dot{z}(t), u(t)$ and $z^2(t)\dot{z}(t)$. The $\dot{z}(t)$ term is one of the terms that contributes the

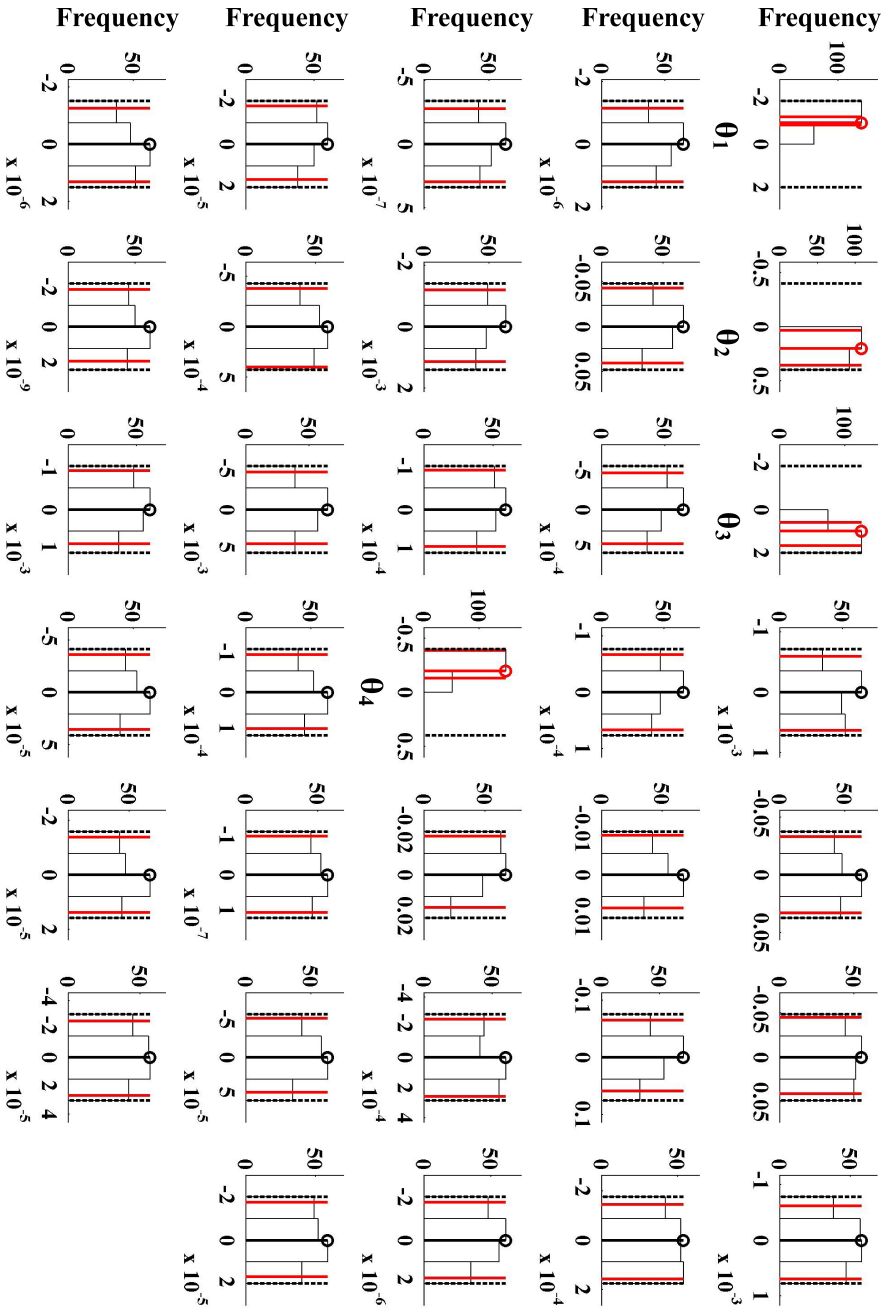


Figure 5.4: Model structure detection using one-stage procedure (Algorithm 5.2: $L = 3$ and $N_s = 200$) for VDPD system. True model terms in red stem are correctly selected, while false model terms in black stem are correctly not selected. The black dotted and red solid vertical lines indicates the prior and quantile values respectively.

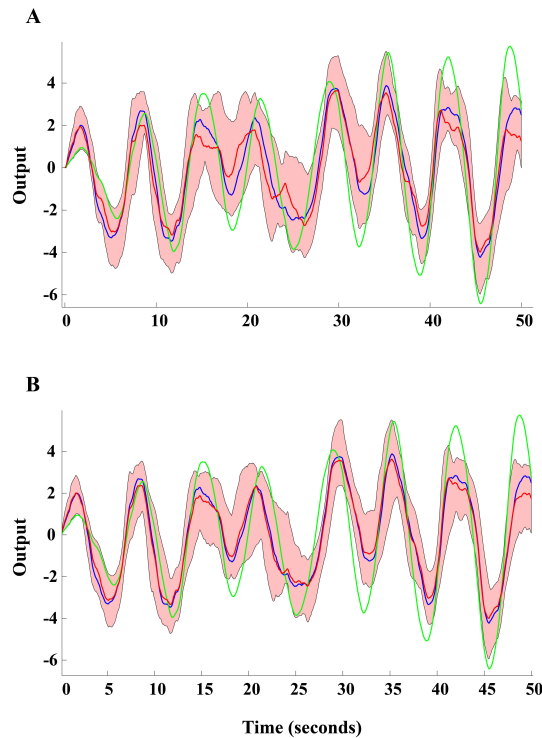


Figure 5.5: Comparison of noise free output (blue), dCTM model (green) and one-stage ABC (red - the shaded region indicates uncertainty from the ABC parameter range). A: one-stage ABC model based on initial parameter estimation (full model set) and B: one-stage ABC model based on re-estimation of parameters (only selected model terms).

least, and from Figure 5.4 it can be seen that it narrowly passes the quantile test, whereas the dCTM approach completely fails to correctly identify it (Table 5.2). The one-stage model structure detection algorithm performs well when the system to be identified has little or no identifiability issues. However, it under-performs when identifiability issues arises (model terms with less sensitive parameters), which the extended two-stage model structure detection algorithm rectifies. This is illustrated in the next subsection.

The improved performance of the one-stage model structure detection algorithm is highlighted by a comparison of simulations (Figure 5.5). It can be seen that the estimates of the parameters are better when only the correct model terms are considered (compare uncertainty - red shaded region in Figure 5.5A and B), this shows the correlative relationship between model terms, that makes the re-estimation of parameters after structure detection essential.

5.6.3 Case 3: model structure detection and derivative order model selection of a test system with identifiability problem

To investigate the performance and accuracy of the proposed two-stage model structure detection algorithm, a test system with identifiability problem is used with increasing measurement noise. Here again, the results were compared to the dCTM approach (note, for the dCTM approach, the tuning parameters for smoothing were chosen as discussed in section 3.7.2).

The test system used was

$$\begin{aligned} \dot{z}(t) = & \theta_1 z(t) + \theta_2 \dot{z}(t) + \theta_3 u(t) + \theta_4 \dot{z}^2(t) \\ & + \theta_5 z(t)u(t) + \theta_6 u^3(t), \end{aligned} \quad (5.18)$$

$$y(t) = z(t) + e(t), \quad (5.19)$$

where $e(t)$ was defined as zero-mean Gaussian random noise with variance (i) $\lambda^2 = 4 \times 10^{-5}$ for SNR of 20dB, (ii) $\lambda^2 = 4 \times 10^{-4}$ for SNR of 10dB. The parameter vector was set to $\theta = (-2, -3, 1, 4, 10, 2.5)$. The excitation signal was set to a zero-mean uniform random sequence in the range $(-1, 1)$ band-limited to 20 Hz. For parameter estimation using ABC-SMC the parameter size was set to $N_s = 200$ and the number of population iterations to $L = 10$. The nonlinear order was set to $q = 3$ and the derivative order to $n_i = 2$ (except for the derivative selection test described below).

The dCTM and one-stage ABC algorithms performed well at high SNR (20dB) with the correct model terms being chosen but worsened with increasing noise levels (SNR=10dB). The two-stage algorithm performed well, however, even at higher noise levels, correctly identifying all terms (Table 5.3). The much improved performance of the two-stage ABC algorithm is highlighted by a comparison of simulations (Figure 5.6E and F).

The one-stage ABC algorithm failed to pick model terms with less sensitive parameters, unexplained variance (relative to the regressed output) of these terms are: $\sim (2.29, 2.22, 2.07)$ that corresponds to $\dot{z}(t)$, $\dot{z}^2(t)$ and $z(t)u(t)$ (Figure 5.6C). However, the $z(t)$ term has a low unexplained variance of 2.20, but its contribution to the dynamics is much more significant thus being correctly picked by the one-stage ABC algorithm. The robustness and efficiency introduced by the Cha-Srihari measure can be seen in Figure 5.6A-B and G. In Figure 5.6G, the terms whose pos-

terior distribution's variance > 1 , captures how much their uniform priors have evolved into their respective posterior distributions (represented as histograms; it is assumed that the more the prior evolves, the greater its significance). The $\dot{z}(t)$ and $z(t)u(t)$ terms with corresponding parameters of θ_2 and θ_5 are demonstrated to have evolved the most in comparison to other model terms shown in Figure 5.6G, which the Cha-Srihari measure have correctly identified (Figure 5.6A), thus granting more importance. This demonstrates that the Cha-Srihari measure is only used for efficiency purpose, as further computation could be avoided if a required threshold (prediction accuracy) is achieved. The $\dot{z}^2(t)$ term with corresponding parameter θ_4 , has a very low significance towards the dynamics, which contributed to its lower ranking according to the Cha-Srihari measure. However, the BIC was still capable of correctly identifying it, which further shows the robustness of the two-stage ABC algorithm.

To demonstrate the selection of the derivative order, two models of derivative order $n_i = (2, 3)$, were obtained using the two-stage ABC algorithm and were compared as explained in section 5.4.3 (using the 10 dB input-output data) and the results seen in Figure 5.6D). From 200 samples, the model with $n_i = 2$ was selected 187 times and the other model with $n_i = 3$ was selected 13 times. The Bayes factor in this case, $B_f(1, 2) = \frac{187}{13} = 14.4$, correctly provided strong evidence in favour of $n_i = 2$ rather than $n_i = 3$, demonstrating the effectiveness of this model selection approach.

5.7 Summary

In this chapter, a computational Bayesian identification framework for nonlinear CT systems that utilises a simulation approach as opposed to a regression approach is developed. The main contribution of this algorithm to the suite of methods available for CT nonlinear system identification is that the signal derivative free approach and the estimation of the model parameter uncertainty by constructing a distribution. The identification algorithm uses the ABC-SMC method, which is a rejection sampling technique for inferring parameters of a model. Parameter distributions intrinsically generated by ABC-SMC estimation algorithm is used to drive term selection by significance testing. The simulation results demonstrate the high fidelity of the ABC approach to increase in noise levels in the measurements.

This developed identification framework will aid the quantification and charac-

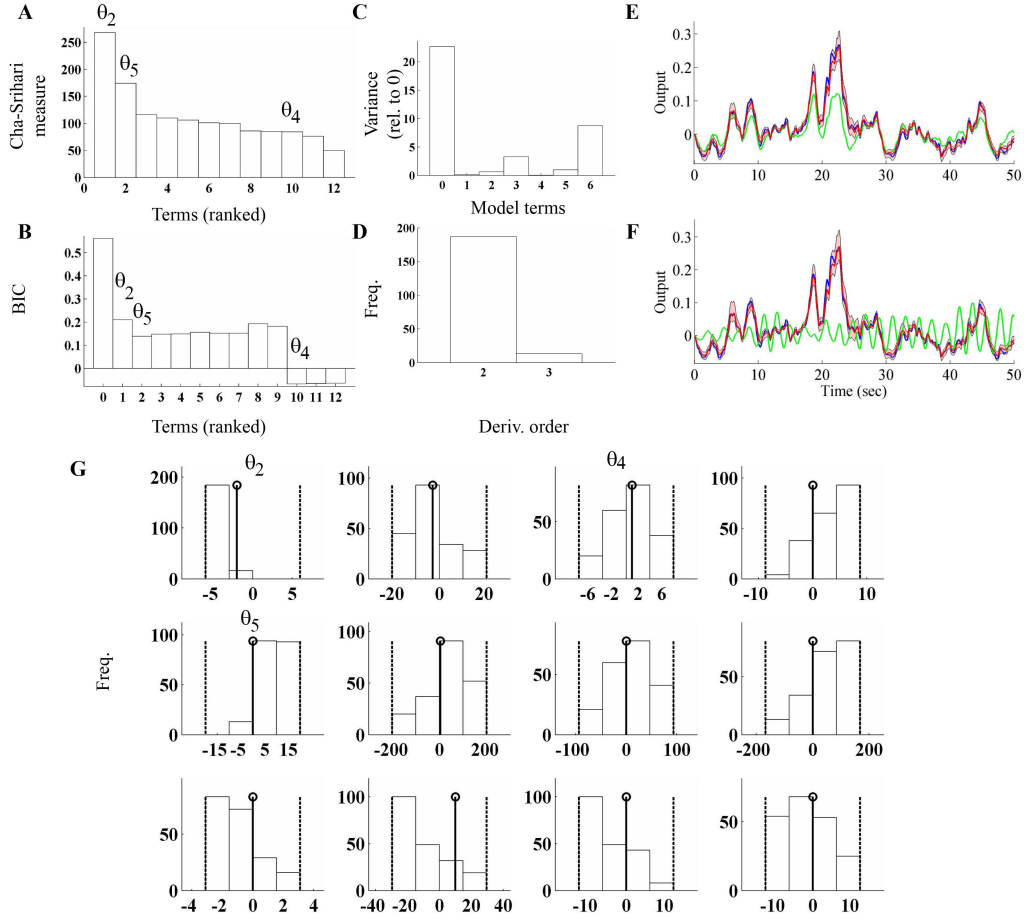


Figure 5.6: Model structure detection using two-stage procedure (Algorithm 5.3: $L = 3$ and $N_s = 200$) and model selection procedure (A-D and F-G shows results for SNR=10dB). A. Cha-Srihari measure of the 12 model terms whose variance > 1 (the correct model terms are indicated), B. The consequent BIC score of models when model terms are added in order of sensitivity (the correct model terms are indicated and the BIC score under 0 represents the model when no terms from pool 2 are added), C. The variance relative to zero of each model term in the system is shown, indicating their contribution to the dynamics (the bar under 0 quantifies that of the system's output), D. The derivative order model selection procedure of model order's $n_i = (2, 3)$, E and F. Comparison of noise free output (blue), dCTM model (green) and two-stage ABC (red - shaded region indicates uncertainty from ABC parameter range): E - SNR = 20dB and F - SNR = 10dB, and G. The posterior distribution (histogram) of the the 12 model terms whose variance > 1 (the correct model terms are indicated).

Table 5.3: Identified models from the test system in case 3. ABC1 refers to the one-stage ABC identification method, and ABC2 refers to the two-stage ABC identification method.

True System	SNR = 20 dB			SNR = 10 dB		
	dCTM	ABC1	ABC2	dCTM	ABC1	ABC2
$-2.00y(t)$	$-2.00y(t)$	$-2.54y(t)$	$-1.79y(t)$	$-9.25y(t)$	$-2.40y(t)$	$-1.76y(t)$
$-3.00\dot{y}(t)$	$-1.25\dot{y}(t)$	$-4.46\dot{y}(t)$	$-2.46\dot{y}(t)$	—	—	$-2.56\dot{y}(t)$
$1.00u(t)$	$0.77u(t)$	$1.12u(t)$	$1.13u(t)$	$0.65u(t)$	$1.14u(t)$	$1.14u(t)$
$4.00\dot{y}^2(t)$	$3.53\dot{y}^2(t)$	—	$3.88\dot{y}^2(t)$	$4.02\dot{y}^2(t)$	—	$4.11\dot{y}^2(t)$
$10.00y(t)u(t)$	$5.09y(t)u(t)$	—	$9.20y(t)u(t)$	—	—	$9.14y(t)u(t)$
$2.50u^3(t)$	$0.52u^3(t)$	$3.77u^3(t)$	$3.42u^3(t)$	—	$3.07u^3(t)$	$3.35u^3(t)$
—	—	—	—	$26.71y^2(t)$	—	—
—	—	—	—	$-0.01u(t)\dot{u}^2(t)$	—	—

terisation of variability in gene expression that is observed in different cell populations, in a principled and rigorous way (as discussed in the earlier chapter). This is demonstrated in Chapter 7 for the experimental data collected in Chapter 6.

Chapter 6

System fabrication, experimental design and data acquisition

6.1 Introduction

In this thesis, the focus is not only based on developing a data-driven modelling framework to derive dynamic models that will enhance the characterisation of synthesised genetic parts. Rather, the aim was to give equal emphasis on data acquisition through designed experimentation, which the developed data-driven modelling framework and analytical tools could be applied to. That will subsequently help in providing answers to key questions related to characterisation of genetic parts, such as the process of characterisation itself. This was made possible by the cross-departmental collaboration with the "ChELSI group" in the department of chemical and biological engineering (<http://www.shef.ac.uk/chelsi>), including other services provided by the University of Sheffield, such as the deoxyribonucleic acid (DNA) sequencing service in the medical school. This collaborative opportunity helped to establish a friendly atmosphere for discussions, knowledge sharing and experimentation related to the research of key fundamental challenges in achieving effective characterisation of genetic parts.

The need to gather more experimental data of the transcriptional regulatory system - BBa_T9002 (http://parts.igem.org/Part:BBa_T9002) was clearly established in Chapter 4. The experimental data should consist of both cell growth and GFP measurements starting from lag phase till death phase. Transcriptional regulatory systems are regarded as one of the most simplistic genetic functional modules which are frequently used to design higher-order genetic parts. The need to characterise commonly used functional modules robustly, in order to simplify and aid

the design of higher-order genetic parts is crucial. Furthermore, the limitation of the data-driven dynamic model derived in Chapter 4 using the narrow time-series experimental data, is its limitation in predicting the dynamics of the BBa_T9002 system beyond the quasi steady state (mid-point in the exponential growth phase). This is a drawback, as output predictions of a genetic part at all stages of the cell growth cannot be attained, which a model of an operational prototype will and should be required to do. Also, the narrow time-series experimental data does not fully capture the compromising effect of the system's cell growth and protein expression on each other. This is taken into account in this thesis and, experimental data are obtained at all stages of cell growth. This imposes a change in the model structure of the data-driven dynamic models as will be shown in Chapter 7 in comparison to the derived model in Chapter 4.

As mentioned in the first paragraph, a key question about the process of characterisation is raised in this thesis. What influence does the reporter cascade, which includes both the ribosome binding site and fluorescence protein, have on the functional module and vice versa, in addition to the overall dynamics of the whole system? In this thesis, the dynamics of a system observed is termed the "relative" dynamics respective to the reporter cascade used. If it does, how much of an influence is it? Is it quantitatively large and if so, is it appropriate to use reporter cascade for the characterisation of genetic parts? For BBa_T9002 system, the functional module is BBa_F2620 (http://parts.igem.org/Part:BBa_F2620) - the receiver cascade. When higher-order genetic parts are designed, different functional modules are usually synthesised and ligated together, normally excluding the reporter cascade. The experimental norm for the process of characterisation is the use of a fluorescence protein to monitor the dynamics of the functional modules. However, little has been done in the literature to verify the influence of a functional module and reporter cascade on each other. There are three conclusions that can be drawn from such a study, which are: (i) the reporter cascade monitors and delivers the exact dynamics exhibited by the functional module (reporter cascade does not have its own dynamics), (ii) the reporter cascade exhibits dynamics of itself which is only being observed in the experimental data or (iii) both the functional module and reporter cascade dynamics have being observed in the experimental data, where dynamics of one dominates the other. It can be argued that option (ii) is an unlikely scenario because there are several different genetic parts (functional modules) which have been designed, and monitored by the same native GFP and exhibit different dynamics (Elowitz and Leibler, 2000, Gardner et al., 2000, Toettcher et al., 2011).

Here, a novel experimentation is devised to answer some of the questions raised in the above. This required the design of a new genetic part. Therefore in this chapter the biofabrication, that exploits the top-down design approach of genetic parts needed for experimentation are discussed. This was made possible by using off-the-shelf functional modules which are readily available in the registry of standard biological parts RSBP. An IGEM laboratory group was created for this purpose - "Wright Lab" (<http://igem.org/Lab.cgi>), allowing the order of the 2012 DNA distribution kit (all genetic parts used in this thesis are from the 2012 distribution kit only). The design of a similar system to the BBa_T9002 was undertaken, in which the functional module - BBa_F2620 was ligated with a different reporter cascade - BBa_J06702 (http://parts.igem.org/Part:BBa_J06702) rather than BBa_E0240 (http://parts.igem.org/Part:BBa_E0240). By keeping the functional module the same, the aim was to investigate if the reporter cascades have an influential effect on the "relative" dynamics of the system, thereby reflecting on some questions raised in the above paragraph.

In this chapter, a summary of the main experimental protocols carried out to accomplish the biofabrication of the genetic parts are detailed. The stage by stage process in which this experimental protocols are applied is also discussed. The experimental setup and procedures for the data acquisition of cell growth and protein expression measurements, for both systems (BBa_F2620 ligated to two different reporter cascades) are outlined. The chapter is concluded by reporting the various data collected from experiments.

6.2 Experimental protocols

The main experimental protocols practiced during the biofabrication of the genetic parts are described here. The experimental protocols were practiced to achieve the: (i) transformation of DNA into host cells, (ii) cloning of DNA samples, (iii) screening of DNA samples, (iv) cutting and ligation DNA sequences, and (v) preparation of growth media. The experimental protocols are mentioned and recalled again in the next section to demonstrate at what stages of the biofabrication process they were applied. Most experimental protocols used are based on the standard approaches, unless stated otherwise in which case a website link would be provided.

6.2.1 Transformation

In order to transform plasmids (with or without DNA inserts) into *E. coli* strains, one of the two protocols could be followed: chemical or electrical transformations. By identifying the fact that the *Escherichia coli* (*E. coli*) strain to be used is either chemically competent (*e.g.* DH5 α strain) or electrically competent (*e.g.* K12 strain), will influence which protocol is to be implemented.

It is also important to identify if the *E. coli* strain is going to be used for cloning or characterisation purposes. Cloning strains are preferred for having reduced nuclease activity, while characterisation strains are preferred for having decreased protease activity.

Chemical transformation

The chemical transformation protocol proceeds as follows:

- Prepare fresh lysogeny broth (LB) plates with ampicillin resistance.
- Streak out chemically competent *E. coli* strain (*e.g.* DH5 α) onto a LB plate. This serves as a control (done in parallel), where no growth of bacteria colonies should be seen overnight, due to the presence of ampicillin resistance. Direction: (i) keep bacteria strain (glycerol stock) in ice at all time, (ii) streak by dipping in a streaking stick onto the glycerol stock and across the plate, and (iii) carry out the process close to the flame.
- Add 2 microlitres (μL) of DNA sample (either obtained from miniprep or maxiprep protocol) to $\sim 10 - 20 \mu\text{L}$ of *E. coli* strain.
- Leave in ice for 30 minutes.
- Leave in floater at 42 degrees Celsius ($^{\circ}\text{C}$) for 90 seconds.
- Leave in ice for 3 minutes.
- Add 150 μL of LB medium into it and place it in a floater inside the incubator at 37 $^{\circ}\text{C}$, 200 revolutions per minute (*rpm*) for 2 hours.
- Plate out all the solution onto a LB plate with ampicillin resistance using a plating stick.

Electrical transformation - electroporation

The electrical transformation protocol proceeds as follows:

- Prepare fresh LB plates (no antibiotics).
- Streak out electrically competent *E. coli* strain (e.g. K12 - MG1655) onto the LB plate and allow it to grow overnight. Direction: (i) keep bacteria strain (glycerol stock) in ice at all time, (ii) streak by dipping in a streaking stick onto the glycerol stock and across the plate, and (iii) carry out the process close to the flame.

On the following day, further procedures outlined below are followed,

- Add 10 millilitres (*mL*) of LB medium into a falcon tube (no antibiotics), slice a single colony of *E. coli* strain from the overnight grown plate and add it to the LB medium in the falcon tube.
- Leave the falcon tube in the incubator at 37°C, 200 *rpm* for overnight.

On the following day, further procedures outlined below are followed,

- Prepare a 1:100 dilution of overnight culture (with no antibiotics) in 10 *mL* of LB medium.
- Leave the diluted culture in the incubator at 37°C, 200 *rpm* for 2 - 3 hours, until an absorbance measurement (growth level - OD600) of 0.4 - 0.7 is achieved.
- Place the falcon tube containing the culture in a centrifuge at 4°C, 4000 *rpm* for 15 minutes. Retrieve the falcon tube and discard the supernatant.
- Resuspend the cell pellet in 1 *mL* of chilled MQ water. Note, MQ water is deionised filtered water.
- Suspend the solution into a 2 μ L eppendorf tube and spin it down using a centrifuge at 4000 *rpm* for 1 minute and then discard the supernatant.
- Repeat the above 2 steps three times.
- Resuspend the sediment in 100 μ L of MQ water and split the suspension into 2 solutions (50 μ L each).

- Add 5 μL of DNA sample into one of the solution and leave it for 30 minutes. Set aside the other solution as a control.
- Aliquot the suspension (with DNA sample) into a chilled curvette. Direction: (i) carefully aliquot it near the membrane (without damaging the membrane) and (ii) carry out the process close to the flame.
- Place the curvette into the electroporation device (Figure 6.1E). Choose appropriate settings (e.g. K12 strain and 2 millivolts) and electroporate.
- Slowly resuspend the solution in the curvette with 1 mL of LB medium and leave the solution in the incubator at 37°C , 200 rpm for 2 hours.
- Spin the solution down in a centrifuge at 6000 rpm for 1 minute.
- Remove 900 μL of the spun down solution, resuspend the rest with 100 μL of LB medium and pour the solution onto a LB plate with ampicillin resistance. Plate it out using a plating stick.

6.2.2 Preparation of M9 supplemented media

This protocol is replicated from

http://openwetware.org/wiki/Endy:M9_media/supplemented. For 1 L of M9 supplemented media, combine the following solutions using a sterile technique:

- 500 mL of $2 \times \text{M9}$ salts.
- 30 mL of 10 milligramme/millilitre (mg/mL) thiamine hydrochloride.
 - Dissolve 10 mg/mL of MQ water.
 - Filter sterilise using a 0.22 micrometre filter.
 - Light-sensitive, therefore store covered.
- 10 mL of 40% glycerol.
- 20 mL of 10% casamino acids.
- 20 mL of 0.1 molar (M) magnesium sulfate.
- 200 μL of 0.5 M calcium chloride.
- 419.8 mL of sterile MQ water.

Some precipitation may be noticed during preparation but precipitate should go back into solution once volume is brought up to 1 L with sterile water. Filter sterilise once the media is made up and store under 4°C .

6.2.3 Colony screening

After each transformation protocol is carried out, colony screening is highly advised. It helps to determine if most bacteria colonies seen growing on the overnight LB plate have the required plasmid insert in them.

- Slice out different colonies from the overnight grown LB plate and: (i) mix it in separate 10 mL of LB medium with ampicillin resistance (100 $\mu\text{g/ml}$ concentration), and (ii) streak out onto separate LB plate with ampicillin resistance.
- Place: (i) each falcon tube containing the diluted culture in an incubator at 37°C, 200 rpm for overnight and (ii) each streaked LB plate into a static incubator at 37°C for overnight.

All the streaked LB plates should have bacteria colonies grown in them the next day, since the LB plates contain ampicillin resistance. If this is not seen, it indicates that the transformation protocol carried out was not successful.

6.2.4 Diagnostic gel

The diagnostic gel protocol is carried out to determine the size of a linear DNA under consideration. The linear DNA has to be linearised (cut both strands of DNA) in order to determine its size. This protocol could be also used to detect the presence of a DNA sample as well. This protocol proceeds as follows:

- Add 0.5 grams (g) of agarose in 50 mL 1 × TAE (tris-base, acetic acid and ethylenediaminetetraacetic acid) and microwave the mixture until all the powdered agarose has dissolved.
- Add 2 μL of ethium bromide into the solution.
- Setup the compact gel kit (Figure 6.1B) using the 20 μL plastic frame.
- Pour the solution into the compact gel kit and let it set for ~ 10 minutes, and then pour in the 1 × TAE buffer.
- Add 5 μL of hyperladder 1 (BIOLINE) to the top well.
- Mix 8 μL of MQ water, 2 μL of 5 × loading buffer and 2 μL of DNA sample.
- Then add the above solution to one of the wells of the compact gel kit and set the gel kit for 80 minutes and 50 milliamperes (mA).

The hyperladder 1 - BIOLINE chart shown in Figure 6.1F, is used to determine the size of a linear DNA sample which is calculated relative to nanogrammes per microlitres. This is done by viewing the gel through the ultraviolet imaging system (Figure 6.1C).

6.2.5 DNA assembly protocols

The three main DNA assembly protocols followed were digestion, dephosphorylation and ligation. The digestion protocol is used to cut linear DNA strand at specific regions (depending on the restriction enzyme used), the dephosphorylation protocol is used to prevent cut plasmid from relinking again, while the ligation protocol is used to link ends of two DNA strands together.

Digestion

Always add with order of reactivity:

- Add 4 μL of MQ water, 1 μL of compatible buffer and 4 μL of DNA sample. In order to determine the appropriate compatible buffer, check the company website where the restriction enzymes were obtained from.
- Add 0.5 μL of each restriction enzymes needed (ice kept) to the solution prepared.
- Spin it down and place it inside the incubator using a floater at 37°C, 200 rpm for 2 hours.

A diagnostic gel protocol is carried out to confirm that the correct segment of the DNA strand is cut or the cut DNA sample is send for sequencing. Sequencing is the determination of the nucleotide order of a given DNA sample.

Dephosphorylation

If the digestion protocol is carried out on a DNA strand which is still linked to the plasmid, then dephosphorylation protocol is done in order to prevent the plasmid from relinking its ends together.

- Add 2 μL of alkaline phosphatase to the DNA sample which is linked to the plasmid (kept in ice).
- Leave for 10 minutes.

Ligation

The ligation calculator can be used to calculate the reactant ratio. Always add with order of reactivity.

- Calculate reactant ratio and add appropriate amount of MQ water, DNA sample, ligase buffer (ice kept) and ligase.
- Leave the solution in 4°C for overnight.

A transformation protocol is carried out the next day, to insert the plasmid into the *E. coli* strain. This is normally done with varying ratio of DNA:strain.

6.2.6 Qiagen toolkits

The Qiagen toolkits were used to provide the following protocols: miniprep, maxiprep and gel extraction. Miniprep is used to purify DNA sample from small volume $\sim 1 - 5mL$ of overnight grown cultures (Qiagen, 2012b). Maxiprep is used to purify DNA sample from large volume $\sim 100mL$ of overnight grown cultures (Qiagen, 2012a). While the gel extraction protocol helps in extracting the digested DNA fragment from the agarose gel (Qiagen, 2010), where the DNA fragment is excised under the ultraviolet illuminator (Figure 6.1D) using a scalpel.

6.3 Design and assembly of genetic parts

The design and assembly of genetic parts used in this thesis are outlined in this section. This includes biofabrication procedures and the resulting sequenced data of the assembled systems. "Finch TV"

(<http://www.geospiza.com/Products/finchtv.shtml>) is used here to view DNA sequence traces, which enables one to identify if the sequenced data is good. "ClustalW2" (<http://www.ebi.ac.uk/Tools/msa/clustalw2/>) is a multi-sequence alignment program, which is used to ensure two or more DNA sequences are identical.

The following were obtained from: (i) genetic parts - IGEM 2012 distribution kit, (ii) restriction enzymes, ligase, buffers and *E. coli* strains - "New England Biolabs" (<https://www.neb.com/>) and (iii) sequencing primers - "life technologies" (<http://www.lifetechnologies.com/uk/en/home.html>).

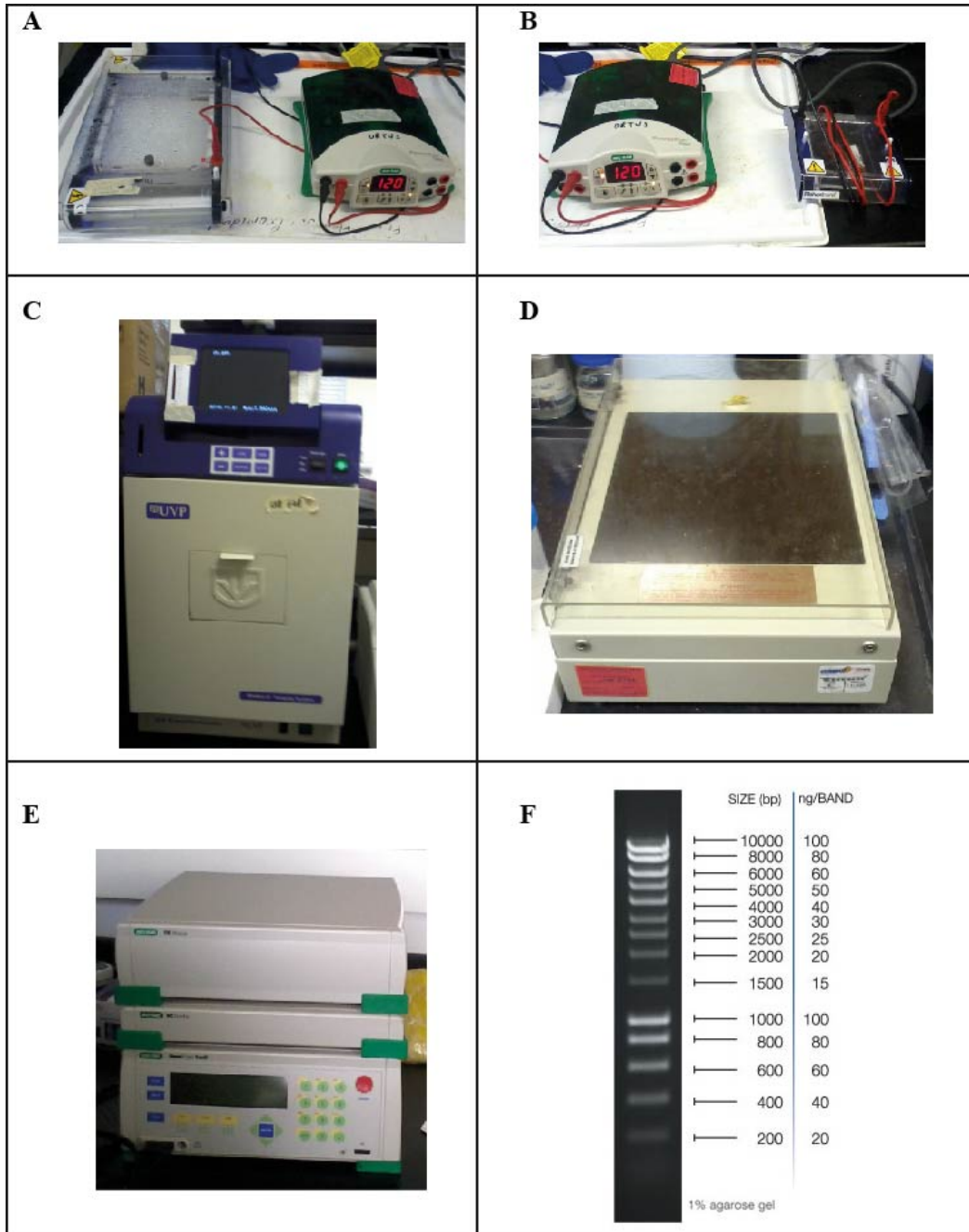


Figure 6.1: Equipments and hyperladder 1 chart used during experimental protocols. A. Large gel kit used for DNA purification or gel extraction, B. Compact gel kit used for diagnostic gel, C. Ultraviolet imaging system, D. Ultraviolet illuminator, E. Electroporation device and F. Hyperladder 1 chart - BIOLINE.

6.3.1 BBa_F2620 and BBa_T9002

The BBa_F2620 and BBa_T9002 genetic parts were retrieved from the IGEM 2012 distribution kit, which were located in kit plate 2 under well 6E and 9A respectively (the pictorial description of both genetic parts can be seen in Figure 4.1). The genetic parts were linked to plasmid backbones "pSB1A2" and "pSB1A3" respectively, that includes a ampicillin resistance tag (Figure 6.2A).

Chemical transformation was carried out for both genetic parts using the *E. coli* strain "XL1 Blue" (for cloning purposes). 10 mL of cultures were grown using the successfully transformed strains. Glycerol stocks and miniprep protocol of the overnight grown cultures were prepared and carried out. Diagnostic gel protocol was carried out to confirm the presence of DNA plasmids in the transformed strains. Using the glycerol stocks, newly streaked LB plates were made to obtain colonies, from which starter cultures were prepared for growing 2L of cultures. The 2L cultures were used to carry out the maxiprep protocol, to obtain high concentrations of BBa_F2620 and BBa_T9002 DNA samples (Figure 6.2B). The DNA samples were sent for sequencing, the traces of the DNA sequence can be seen in Figure 6.2C and D, which shows clearly defined bands indicating good and clear sequence. ClustalW2 was used to validate the sequenced data obtained for both BBa_F2620 and BBa_T9002 genetic parts against their actual sequences which are readily available in RSBP, the comparison showed $\sim 70\% - 80\%$ match with automated analysis for both genetic parts and 100% match for targeted regions (includes sequence obtained from both forward and reverse primers). The DNA samples obtained from the maxiprep protocol were transformed into the *E. coli* strain "K12 MG1655" (for characterisation purposes), in order to preserve the consistency from Chapter 4 and (Canton et al., 2008), from which the experimental procedures are replicated (see section 6.4 for more details).

6.3.2 BBa_J06702 and newly designed genetic part

The BBa_J06702 genetic part is used here as an alternative reporter cascade, which consist of a ribosome binding site - BBa_B0034, monomeric red fluorescence protein (RFP) - BBa_J06504 (also known as "mCherry") (Shaner et al., 2004) and terminator - BBa_B0015 (Figure 6.3). It was retrieved from kit plate 2 under well 8E and was linked to the plasmid backbone "pSB1A2".

Chemical transformation was carried out for BBa_J06702 using the *E. coli* strain "DH5 α " (for cloning purposes). 10 mL of culture was grown using the successfully

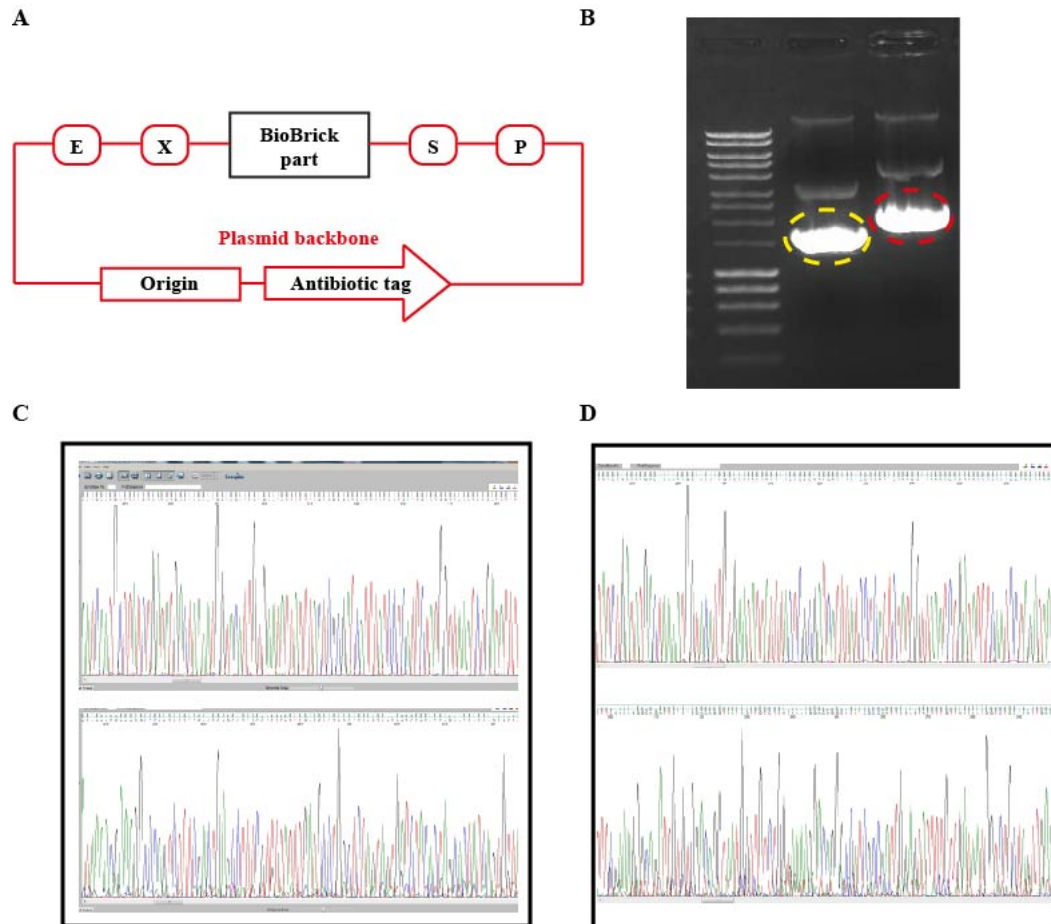


Figure 6.2: The BBA_F2620 and BBA_T9002 constructs. A. The standard plasmid construct (red) of genetic parts in RSBP, where E, X, S and P stands for the restriction sites EcoRI, XbaI, SpeI and PstI respectively, and the antibiotic tag is usually ampicillin, B. The DNA samples obtained from maxiprep protocol is viewed under the ultraviolet imaging system, BBA_F2620 (1061 base pairs) - yellow circle and BBA_T9002 (1945 base pairs) - red circle, C. and D. Snapshots from FinchTV showing the DNA sequence traces of BBA_F2620 and BBA_T9002 respectively (top window showcases the sequence obtained using forward primer, while the bottom window showcases the sequence obtained using the reverse primer).

transformed strain. Glycerol stock and miniprep protocol of the overnight grown culture was prepared and carried out. Diagnostic gel protocol was carried out to confirm the presence of DNA plasmid in the transformed strain. Using the glycerol stock, newly streaked LB plate was made to obtain colonies, from which a starter culture was prepared for growing 2L of culture. The 2L culture was used to carry out the maxiprep protocol, to obtain high concentration of BBa_J06702 DNA sample (Figure 6.4A). The DNA sample was sent for sequencing, the traces of the DNA sequence can be seen in Figure 6.4B, which shows clearly defined bands indicating good and clear sequence. ClustalW2 was used to validate the sequenced data obtained for BBa_J06702 genetic part against its actual sequence which is readily available in RSBP, the comparison showed $\sim 97\% - 99\%$ match and 100% match for targeted regions (includes sequence obtained from both forward and reverse primers).

In constructing the new genetic part, the digestion protocol was carried out on both BBa_F2620 and BBa_J06702, using restriction enzymes SpeI and PstI, and XbaI and PstI respectively. Dephosphorylation protocol was carried out on the double digested BBa_F2620 genetic part, to avoid linkage of the plasmid. A large diagnostic gel was prepared to view the digested genetic parts BBa_F2620 and BBa_J06702, under the ultraviolet imaging system (Figure 6.4C). Using the ultraviolet illuminator, the digested DNA samples were excised (Figure 6.4D) and the gel extraction protocol was carried out. On extracting the digested DNA samples, the ligation protocol was carried out to link the two digested DNA samples, BBa_F2620 with BBa_J06702, which resulted to the new genetic part that is labeled "F2620-RC2" in this thesis (Figure 6.4E). In Figure 6.3, the assembly of "F2620-RC2" system is illustrated graphically. The new genetic part was then transformed into the *E. coli* strain "DH5 α " (for cloning purposes), subsequently permitting the maxiprep protocol to be carried out, in order to obtain high concentration of "F2620-RC2" DNA sample. The DNA sample was sent for sequencing using only a reverse primer, as this should retrace the sequence of BBa_J06702 genetic part. ClustalW2 was used to validate the sequenced data with the comparison showing $\sim 80\% - 85\%$ match and 100% match for targeted regions. The "F2620-RC2" DNA sample was transformed into the *E. coli* strain "K12 MG1655" (for characterisation purposes) to be consistent with Chapter 4 and (Canton et al., 2008), from which the experimental procedures are replicated.

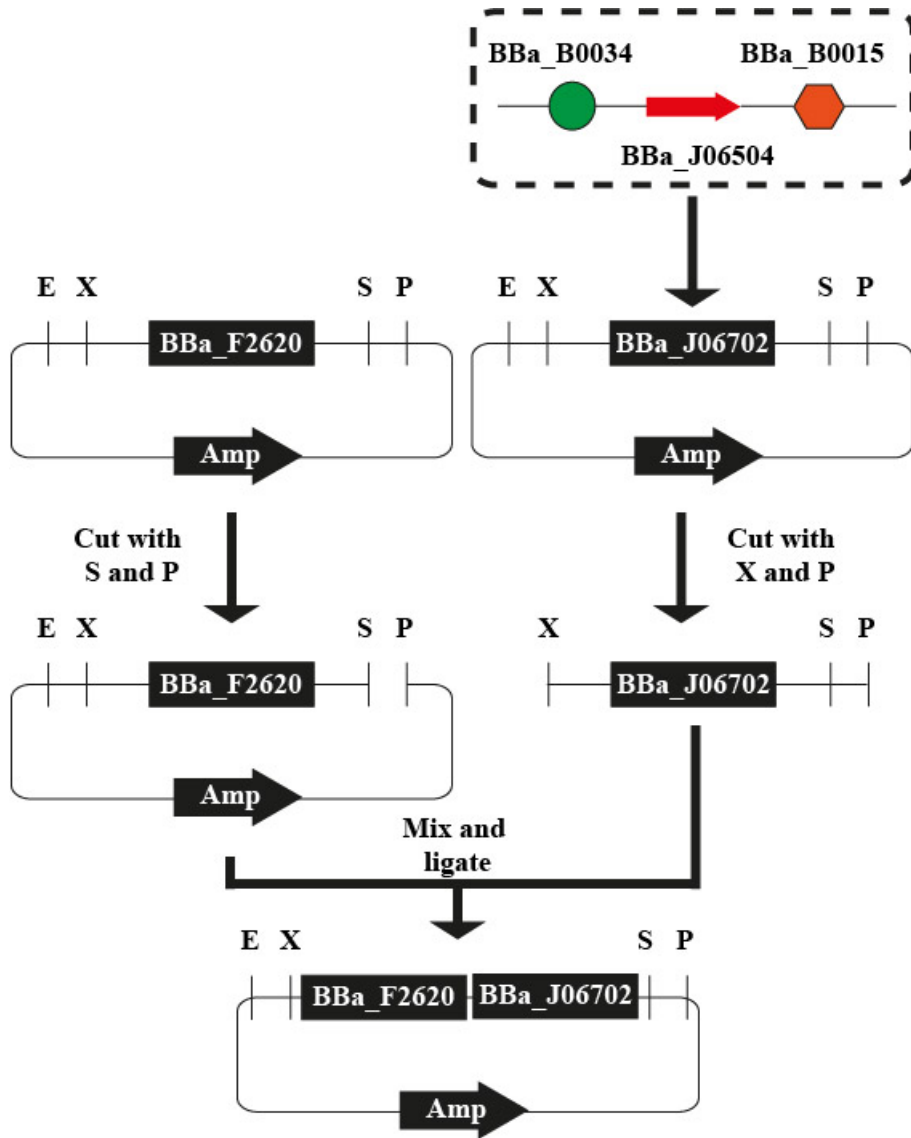


Figure 6.3: The design steps and construct of the new genetic part - "F2620-RC2" system. At top, the pictorial description of the BBa_J06702 genetic part is shown. Below, the design steps used in constructing the new genetic part is shown, where the restriction sites X - XbaI and S - SpeI have similar sequences thereby complimenting each other.

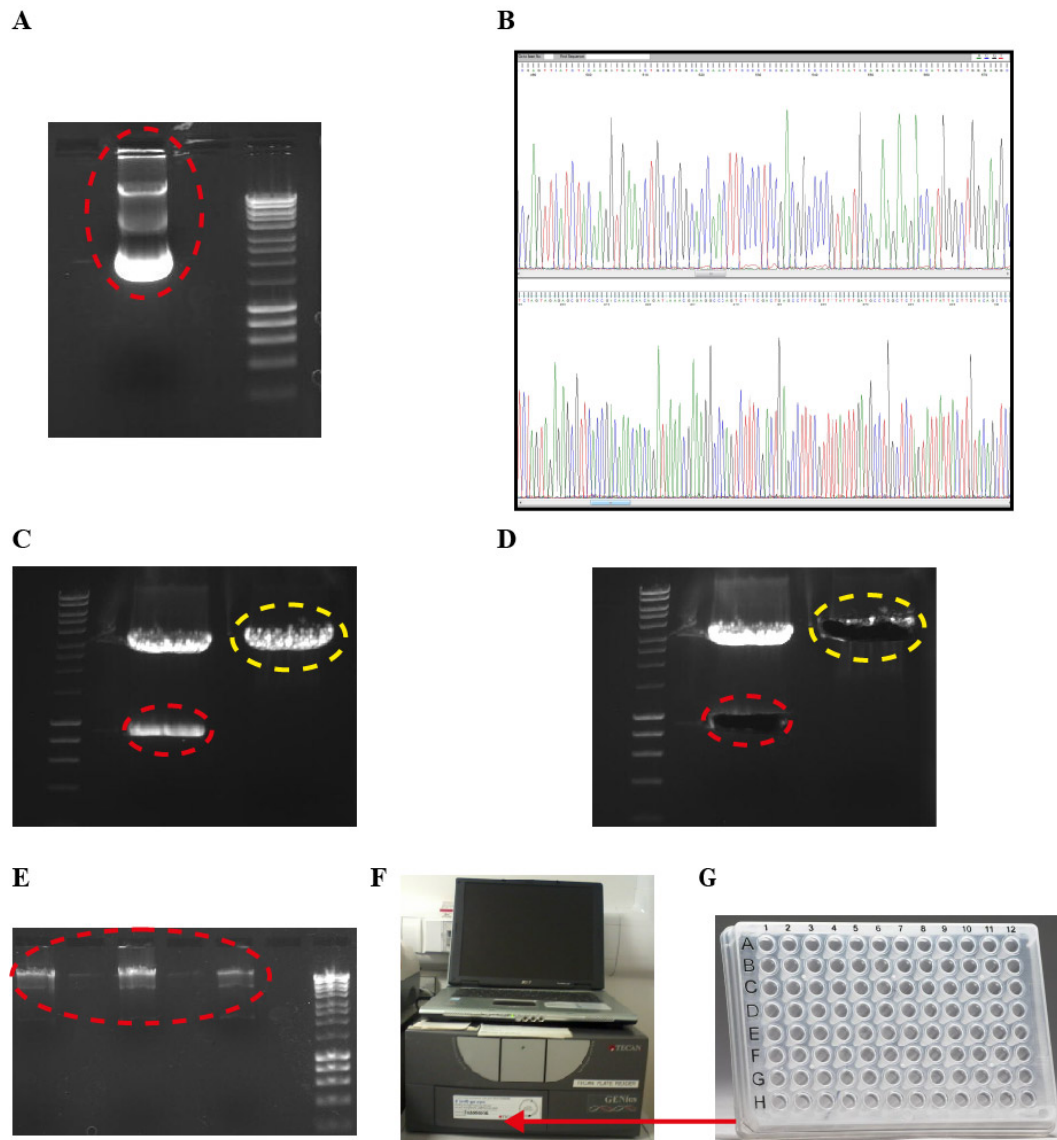


Figure 6.4: Gel imaging, sequencing traces of BBa_J06702 and TECAN GENios microplate reader. A. The DNA sample obtained from maxiprep protocol is viewed under the ultraviolet imaging system, BBa_J06702 (869 base pairs) - red circle, B. Snapshots from FinchTV showing the DNA sequence traces of BBa_J06702 (top window showcases the sequence obtained using forward primer while the bottom window showcases the sequence obtained using the reverse primer), C. The double digested DNA samples of BBa_J06702 (red) and BBa_F2620 (yellow), D. The gel fragments excised for carrying out gel extraction, E. The DNA sample obtained from miniprep protocol is viewed under the ultraviolet imaging system, "F2620-RC2" system (1930 base pairs) - red circle, F. The TECAN GENios microplate reader and G. A 96 well plate prototype.

6.4 Experimental setup and data acquisition procedures

In this section the experimental setup and procedures carried out to obtain the required experimental data are discussed. One should note that the experimental setup and procedures are replicated from (Canton et al., 2008), providing a guiding platform, whose experimental data was modelled in Chapter 4. The alterations imposed here are: (i) an extra genetic part - "F2620-RC2" and (ii) simultaneous measure of cell growth and fluorescence protein in a longer time period of experimentation (lag phase to death phase).

Both BBa_T9002 and "F2620-RC2" systems are transformed into *E. coli* strain - "K12 MG1655", for characterisation purposes. The chemical induction input - 3-oxohexanoyl-L-homoserine lactone (3OC₆HSL) was obtained from Sigma-Aldrich. It was dissolved in dimethyl sulfoxide (DMSO) to a stock concentration of 23.55 mM. Prior to each experiment, the stock concentration was diluted in MQ water to obtain fresh solutions ranging in concentration from 1e-9 M to 1e-3 M. The TECAN GENios microplate reader (Figure 6.4F) was used to measure the cell growth and fluorescence protein measurements, where the compatible software "Magellan" was used to program and control the experimentation.

The experimental procedures are as follows:

- 4 × 50 mL falcon tubes containing 10 mL of M9 supplemented media and 10 μL of 100 mg/mL ampicillin each were prepared.
- 3 single colonies of "K12 MG1655" containing BBa_T9002 or "F2620-RC2" and a single colony of "K12 MG1655" containing BBa_F2620 were inoculated into separate falcon tubes.
- The falcon tubes were placed in the incubator at 37°C, 200 rpm for 15 hours (overnight).
- The cultures were diluted 1:1000 (×2) into fresh 10 mL of M9 supplemented media and 10 μL of 100 mg/mL ampicillin. One batch was placed into the incubator at 37°C, 200 rpm for 4.5 hours (until an OD₆₀₀ of 0.15 is attained). The other batch was reserved (kept in 4°C) and placed into the incubator 15 hours prior the experimentation the next day (under the same condition).
- In eppendorf tubes, appropriate quantities of cultures and 3OC₆HSL in diluted solutions were transferred and mixed. This yielded 8 different final

concentrations - (0, 1e-10, 1e-9, 1e-8, 1e-7, 1e-6, 1e-5 and 1e-4 M) for each culture.

- The mixed solutions in the eppendorf tubes were transferred into a flat-bottom 96 well plate. Three replicate wells were filled for each concentrations of 3OC₆HSL of each culture.
- The plate was incubated in the TECAN GENios microplate reader at 37°C and assayed with an automatically repeating protocol of absorbance measurements (595 nanometres (*nm*)) and shaking (orbital and normal speed for 120 seconds, shake settle time between cycles of 5 seconds). Kinetic time interval between repeated measurements was 178 seconds.

On the following day, after approximately 15 hours, further procedures outlined below are followed,

- Using the other batch (which was reserved), cultures were diluted 1:1000 into fresh 10 mL of M9 supplemented media and 10 μ L of 100 mg/mL ampicillin. They were placed into the incubator at 37°C, 200 rpm for 4.5 hours (until an OD₆₀₀ of 0.15 is attained).
- In eppendorf tubes, appropriate quantities of cultures and 3OC₆HSL in diluted solutions were transferred and mixed. This yielded 8 different final concentrations - (0, 1e-10, 1e-9, 1e-8, 1e-7, 1e-6, 1e-5 and 1e-4 M) for each culture.
- The mixed solutions in the eppendorf tubes were transferred into a flat-bottom 96 well plate. Three replicate wells were filled for each concentrations of 3OC₆HSL of each culture.
- The plate was incubated in the TECAN GENios microplate reader at 37°C and assayed with an automatically repeating protocol of fluorescence measurements (BBa_T9002: 485 *nm* excitation filter, 535 *nm* emission filter or "F2620-RC2": 530 *nm* excitation filter, 610 *nm* emission filter) and shaking (orbital and normal speed for 120 seconds, shake settle time between cycles of 5 seconds). Kinetic time interval between repeated measurements was 145 seconds.

6.5 Overview of experimentally obtained datasets

The Figures 6.5 and 6.6, shows the experimental data collected for systems BBa_T9002 and "F2620-RC2" respectively.

6.6 Summary

In this chapter, the experimental protocols needed to carry out the biofabrication process of the systems BBa_T9002 and "F2620-RC2" are outlined. The experimental setup and procedures used in acquiring the required experimental data are also outlined. Additionally, pictures of gel imaging, validation of sequenced data and reporting of the collected experimental data is presented as part of the wet laboratory work undertaken.

The main reasons for further experimentation was: (i) to collect both cell growth and fluorescence measurements for longer time period (lag phase to death phase), in order to capture and model the full range of dynamics exhibited by a transcriptional regulatory system and (ii) to implement a novel experimentation by assembling a new genetic part "F2620-RC2", which will help to investigate if a reporter cascade has an influential effect on the "relative" dynamics of the system it has been linked to.

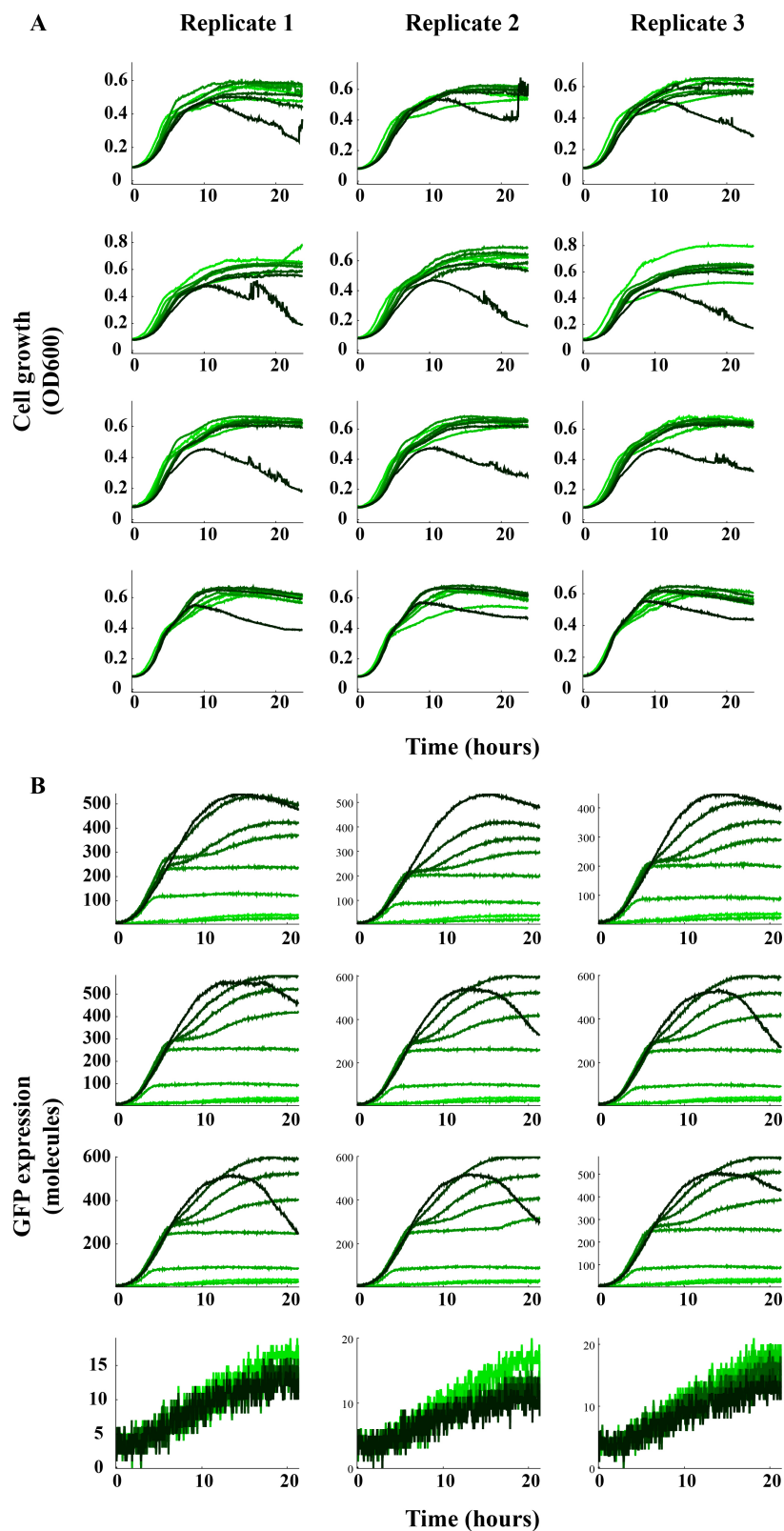


Figure 6.5: Experimental data of BBa_T9002. A. and B. First row - BBa_T9002 colony 1, second row - BBa_T9002 colony 2, third row - BBa_T9002 colony 3, and fourth row - BBa_F2620 colony 1 (control). Response due to 0 M of 3OC₆HSL in light green while response due to 1e-4 M of 3OC₆HSL in dark green.

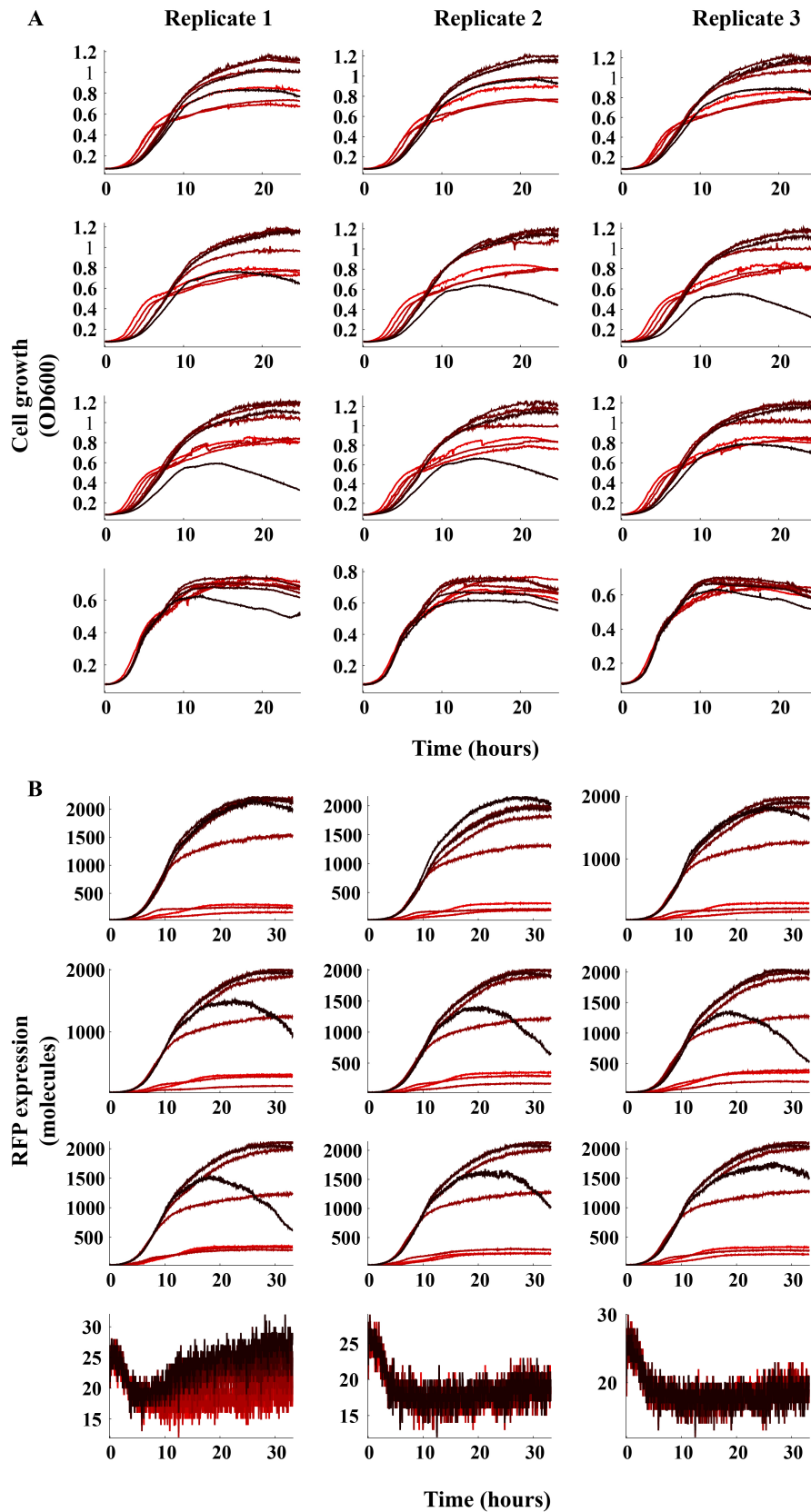


Figure 6.6: Experimental data of "F2620-RC2". A. and B. First row - "F2620-RC2" colony 1, second row - "F2620-RC2" colony 2, third row - "F2620-RC2" colony 3, and fourth row - BBa_F2620 colony 1 (control). Response due to 0 M of 3OC₆HSL in light red while response due to 1e-4 M of 3OC₆HSL in dark red.

Chapter 7

Interpretation of a gene reporter signal and key dynamic design properties

7.1 Introduction

The experimental data collected for both BBa_T9002 and "F2620-RC2" systems were presented in the last chapter. The experimental data of each system consisted of both cell growth and protein expression measurements over time, from lag phase to death phase. The collected experimental data, will allow the developed computational Bayesian identification framework in Chapter 5, to capture and robustly characterise the dynamic properties of the transcriptional regulatory systems - BBa_T9002 and "F2620-RC2". A novel experimentation is devised, where a new transcriptional regulatory system is built - "F2620-RC2" system, which allows one to investigate if a reporter cascade has an influential effect on the "relative" dynamics of the system it has been linked to. The "F2620-RC2" system shares the same functional module - BBa_F2620 as the BBa_T9002 system, however, has a different reporter cascade - BBa_J06702. The "relative" dynamics is the observed dynamics of the whole system with respect to the reporter cascade used.

The devised investigation could result in any one of three likely outcomes: (i) the reporter cascade monitors and delivers the exact dynamics exhibited by the functional module, reporter cascade does not have its own dynamics, (ii) the reporter cascade exhibits dynamics of itself which is only being observed in the experimental data or (iii) both the functional module and reporter cascade dynamics have been observed in the experimental data, where dynamics of one dominates

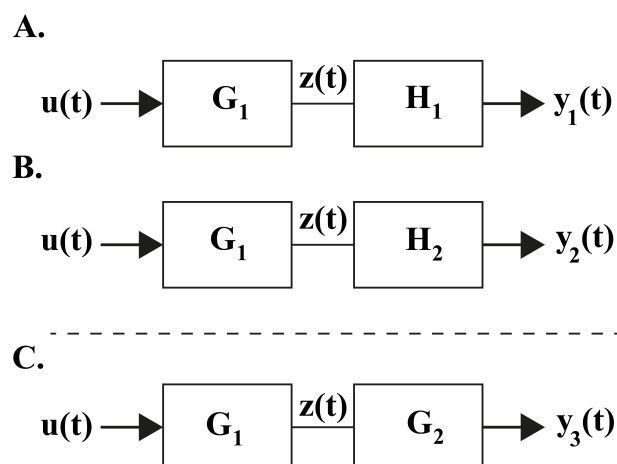


Figure 7.1: Pictorial description of the investigation. The G blocks and H blocks represent functional modules and reporter cascades respectively. In this investigation G_1 - BBa_F2620, H_1 - BBa_E0240, H_2 - BBa_J06702 and G_2 - arbitrary functional module. A. and B. is implemented to achieve the characterisation of G_1 relative to H_1 and H_2 respectively, while C. demonstrates the assembly of an arbitrary higher-order genetic part.

the other. This investigation proves to be very important because the experimental norm in the process of characterisation of a functional module, is the use of reporter cascade to indirectly (relatively) monitor the dynamics of the functional module. However, when higher-order genetic parts are designed, different functional modules are usually synthesised and ligated together, while excluding the reporter cascade. Therefore, if the reporter cascade has an influential effect on the "relative" dynamics of the characterised system, it would be practically incorrect to use the characterised properties of the system as that of the functional module, which would hinder the design of higher-order genetic parts involving the supposedly characterised functional module. The wider question to the synthetic biology community is, is it appropriate to use reporter cascade for the characterisation of genetic parts? This chapter tries to provide the answer to this investigation, where Figure 7.1 gives a pictorial description of the investigation.

Some recent research articles in synthetic biology described below, have reported some shortcomings in the field and their attempts to provide solutions to the shortcomings. Firstly, Ellis et al. (2009) demonstrated the design and assembly of genetic networks (one-order promoter systems) by developing: (i) a library of components (promoters) and (ii) in-silico modelling techniques. This enabled the design and assembly of genetic networks through computational tools, which the early literature in synthetic biology lacked. This was proposed to overcome the

challenge of engineering genetic networks from modular components, which are hampered by: (i) lack of suitable components, while available components tend to have the required functionality but not the quantitative properties needed for design and (ii) extensive post-hoc tweaking of already assembled genetic networks. The library of components developed, consisted of synthesised regulatory promoters, built by varying the sequence of the promoters in a systematic way to achieve diverse promoter strengths. In implementing in-silico modelling, the Hill equation was used to characterise the promoter's strength quantitatively, aiding the design of feed-forward loop networks. By characterising the promoter's strength quantitatively, in contrast to labelling or stating promoters to be either weak or strong, is hugely beneficial for model-based design. However, due to the limitation posed by the Hill equation (only static analysis), as discussed in Krishnanathan et al. (2012) and Chapter 4, certain design specifications of the genetic network's dynamics cannot be achieved, which could be feasible through dynamic analysis. This is daunting, as the few applicable genetic parts are inappropriately characterised, hindering the design of more complex functionalities. Ellis et al. (2009), also reported that the initial characterisation of the promoter's strength, was not adequate enough to aid the design of a two-order promoter system. This was resolved by acquiring extra experimental data from the two-order promoter system, which was used to relatively calibrate the characteristic properties of the one-order promoter system, in order to aid design. This approach is hugely beneficial and further discussed later, by revealing how this approach could be incorporated into the work presented in this chapter.

Daniel et al. (2013) proposed the use of analogue computation in genetic parts rather than digital computation, in order to achieve sophisticated functionalities. Digital computation refers to the use of the two stable states in a bimodal system, the "ON" and "OFF" states when viewed on the logarithmic scale, to perform functions. Whereas, the linear transition (in logarithmic scale) between the "ON" and "OFF" states of a bimodal system, is used to implement analogue computation to perform functions. Daniel et al. (2013), designed a positive feedback mechanism to a transcriptional regulatory system, which provided a wider linear transition range in comparison to that of an open-loop mechanism, thereby allowing for a much wider input range to be used in perturbing the system. This paves way for designing more complex functionalities. The diverse functionalities achieved in Daniel et al. (2013), resulted in diverse static forms, which were modelled either using Hill equation or logarithmic functions. It is not at all surprising that the Hill equation fails to capture non-sigmoidal static forms, which makes the data-driven

identification framework developed in Chapter 5 even more attractive for characterisation, as it is capable of capturing both the static and dynamic properties of differing time responses. Daniel et al. (2013), proposed the use of analogue computation, in order to build more efficient functionalities in environments of cells, which have limitation in cellular resources. The limitation of cellular resources in cells alters the performance of the genetic parts in them, which is usually ignored or not investigated in the design of higher-order genetic parts. Here, the cellular burden introduced by genetic parts on the cells are explicitly quantified, in order to shed some light on how limitations in cellular resources affect the performance of the biological systems.

Breitling et al. (2013) puts forward the need to develop principled approaches to obtain parameter uncertainty of dynamic models used in characterising biological systems, as biological systems will always operate in noisy environments. This occurs because assay conditions vary in comparison to the in-vivo conditions the biological systems operate in and the enzymes in synthesised biological systems have to function in a new cellular environment. Therefore, incorporating the uncertainty of model parameters early on in the design stages will aid predictions that come with specified confidence intervals, which could guide robust designs of genetic parts. In this chapter, the computational Bayesian identification framework developed in Chapter 5 would be implemented to achieve this goal.

The novel experimentation devised in this chapter will enable the investigation of whether the functional module - BBa_F2620, can be robustly characterised regardless that its dynamics not being directly monitored. This would allow one to determine if reporter cascades are appropriate for characterisation of genetic parts. In doing so, the following is achieved in this chapter: (i) "single-cell" models are derived, to characterise the "single-cell" protein expression of the systems - BBa_T9002 and "F2620-RC2", which are later transformed into frequency functions for spectral analysis, (ii) parameters of the cell growth models (for both systems) are estimated, in order to explicitly quantify cellular burden and (iii) population heterogeneity observed in the collected experimental data are characterised, using the parameter distributions naturally generated by the identification framework, thereby predicting the "population-level" protein expression of both systems in the process.

7.2 Biological differences in the reporter cascades

A reporter cascade is made up of a ribosome binding site (RBS) and a fluorescence protein. The biological differences in the reporter cascades - BBa_E0240 and BBa_J06702 of the systems - BBa_T9002 and "F2620-RC2" respectively, are summarised in Table 7.1. The reporter cascade - BBa_E0240 is made up of the RBS - BBa_B0032 and green fluorescence protein (GFP), whereas the reporter cascade - BBa_J06702 is made up of the RBS - BBa_B0034 and red fluorescence protein (RFP).

The fluorescence proteins, GFP and RFP have very similar properties. However, the GFP is a weak dimer, whereas the RFP is a monomer. The effect of this property on the "relative" dynamics of the systems they are linked to, are not known.

The RBS - BBa_B0032 of BBa_T9002 system, is shown to have less relative strength compared to that of the RBS - BBa_B0034 of "F2620-RC2" system. From the literature, relatively strong RBS are proven to aid higher protein expression (Salis et al., 2009). However, this cannot be verified here because the protein expressions of the two systems are monitored using two different fluorescence proteins. Theoretically, the cellular burden on the cells by the "F2620-RC2" system should be greater than that of the BBa_T9002 system, since more proteins are expected to be expressed by the "F2620-RC2" system.

7.3 Modelling BBa_T9002 and "F2620-RC2" systems

Using the collected experimental data for both BBa_T9002 and "F2620-RC2" systems, the aim is to derive data-driven nonlinear continuous-time (CT) dynamic models to capture and characterise the "population-level" protein expression properties of both systems. The novel identification framework which uses a computational Bayesian approach, demonstrated in Chapter 5 is implemented for this purpose. The two additional advantages of the novel identification framework used here are: (i) parameter distributions are naturally generated, giving the user a clear description of uncertainty and (ii) the method is well suited to noisy signals because it avoids the need to estimate signal derivatives. For the identification of the systems under this framework, the input and output signals are defined as 3OC₆HSL concentration and protein expression respectively, where protein expression refers to either GFP or RFP for BBa_T9002 or "F2620-RC2" systems respectively. The protein expression is preferred as the output signal due to its stable nature, which can be seen in Figure 6.5B and 6.6B.

Table 7.1: Differences in the reporter cascades. The properties with superscript indicated as *i* or *ii* are found from RSBP or (Shaner et al., 2005) respectively.

BBa_E0240		BBa_J06702	
RBS (BBa_B0032)	GFP(GFPmut3b)	RBS (BBa_B0034)	RFP (mCherry)
Relative RBS strength ^{<i>i</i>} - 31%	Brightness ^{<i>ii</i>} - $39(nM * cm)^{-1}$	Relative RBS strength ^{<i>i</i>} - 100%	Brightness ^{<i>ii</i>} - $16(nM * cm)^{-1}$
	Photostability ^{<i>ii</i>} - 174 seconds		Photostability ^{<i>ii</i>} - 96 seconds
	Oligomerisation ^{<i>ii</i>} - weak dimer		Oligomerisation ^{<i>ii</i>} - monomer
	Excitation wave-length ^{<i>i</i>} - 501nm		Excitation wave-length ^{<i>i</i>} - 587nm
	Emission wave-length ^{<i>i</i>} - 511nm		Emission wave-length ^{<i>i</i>} - 610nm
	Maturation time ^{<i>ii</i>} - 8 minutes		Maturation time ^{<i>ii</i>} - 15 minutes
	DNA length ^{<i>i</i>} - 720 base pairs		DNA length ^{<i>i</i>} - 714 base pairs

7.3.1 Experimental data

The experimental datasets used here are shown in Figure 6.5 (for BBa_T9002 system) and 6.6 (for "F2620-RC2" system). The experimental dataset of each system, consisted of time-series recording of cell growth and protein expression (GFP - BBa_T9002 system and RFP - "F2620-RC2"), which were observed over approximately 20 hours for BBa_T9002 system and 30 hours for "F2620-RC2" system, over 8 different 3OC₆HSL input concentrations: 0, 1e-10, 1e-9, 1e-8, 1e-7, 1e-6, 1e-5, 1e-4 molar (*M*). The experimental dataset of each system were analysed from 3 colonies of the respective system. There were 3 replicates for each colony, resulting in 9 experiments in total that are analysed for each system. To summarise, there are 2 experimental datasets, each dataset has 9 experiments and each experiment consists of cell growth and protein expression measurements over 8 different 3OC₆HSL input concentrations.

The assumption of a constant input level over time is taken here (same as Chapter 4). Accordingly, the input signal is assumed to be constant which is equivalent to the initial concentration of 3OC₆HSL, for the full duration of the experimentation (from the time of induction to death phase). Bioassay is a popular procedure to measure the concentration of 3OC₆HSL molecules over time, however, the measurement could turn out to be insignificant because the concentration levels are very low and the bioassay procedure itself is wasteful. Therefore the following hypothesis is suggested, the 3OC₆HSL molecules disintegrate and become available again after each complex formation with LuxR and transcriptional activation. However, 3OC₆HSL molecules are likely to degrade over time and the degradation is assumed to be of very small concentration.

7.3.2 Model representation

The primary aim in the identification problem is to capture and characterise the "population-level" protein expression properties of both systems, however, cell growth and "single-cell" protein expression of the systems were required to be modelled, in order to achieve the primary aim. These insights obtained from the experimental data, suggests a different representation of both systems.

The cell growth measurements over 8 different 3OC₆HSL input concentrations are different over time (Figure 6.5A and 6.6A), in each experiments in both experimental datasets corresponding to BBa_T9002 and "F2620-RC2" systems. The number of cells, which is modelled using a cell growth model, influences the "population-

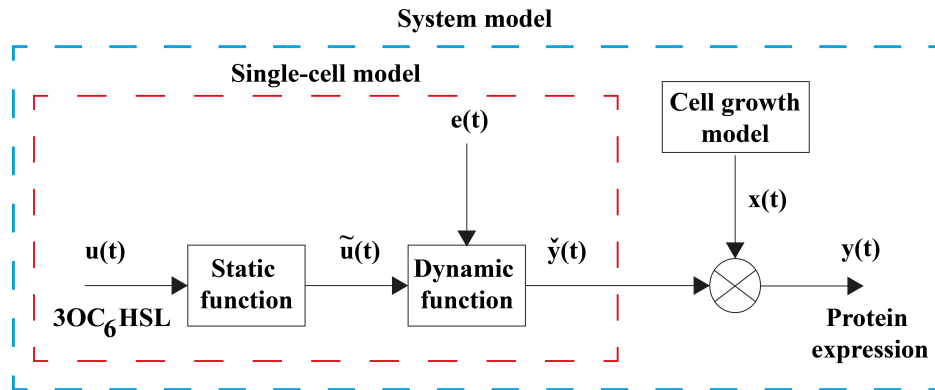


Figure 7.2: System representation. The model structure representation of the system model - blue dashed box, where the dynamic function (data-driven) corresponds to the CT-NARX model, the cell growth model corresponds to the modified Lin's model and the "single-cell" model is represented in the red dashed box.

level" protein expression of the systems.

It was vital to capture the relationship between the input and output signals only, to represent the protein expression dynamics of the systems, in order to aid spectral analysis which is discussed later. This prompted the derivation of models at the "single-cell" level using only the input and output data to characterise the protein expression dynamics of the systems over different induction levels of 3OC₆HSL. The "single-cell" model describes the average number of proteins expressed in each cell, hence, the input and output signals to the "single-cell" model are defined as 3OC₆HSL concentration and average protein expression respectively. The "relative" dynamics ("population-level" protein expression) of the systems are obtained by multiplying the average number of proteins expressed in a single cell by the number of cells present in each well (cell population) as shown in Figure 7.2.

The "single-cell" protein expression responses of both BBa_T9002 and "F2620-RC2" systems over 8 different 3OC₆HSL input concentrations, approximately overlays each other when normalised with respect to their corresponding responses (to remove the static gain effects). This is a feature which is well captured by cascade models. Therefore, similar to Chapter 4, the "single-cell" protein expression response for both BBa_T9002 or "F2620-RC2" systems are described by a single static and dynamic function by taking inspiration from the structure of a Hammerstein model.

The dynamic function is approximated by a CT nonlinear autoregressive model with exogenous input (NARX), which captures the deterministic part of the underlying "single-cell" protein expression dynamics of the systems. The noise term $e(t)$ is assumed to capture the process noise - intrinsic noise which arises due to cell to cell differences.

Cell growth model

In Lin et al. (2000), a cell growth model (Lin's model) is developed based on the time dependent changes of growth rate $\mu(t)$, which is able to predict the lag, exponential and stationary phase of a microbial culture,

$$\mu(t) = \mu_{max} \left(\frac{1}{1 + e^{-k_{in}(t-t_{in})}} \right) \left(\frac{1}{1 + e^{k_{de}(t-t_{de})}} \right), \quad (7.1)$$

$$\dot{x}(t) = \mu(t)x(t), \quad (7.2)$$

where the maximum increasing rate of $\mu(t)$ is k_{in} , the maximum decreasing rate of $\mu(t)$ is k_{de} , the time point when the increasing rate of $\mu(t)$ equals k_{in} is t_{in} and the time point when the decreasing rate of $\mu(t)$ equals k_{de} is t_{de} .

Here, the Lin's model is modified in order to capture the decay phase in a microbial growth as well, by including an exponential decay term. For either systems, a single modified Lin's model, with identified parameters and varying initial conditions, is used to predict its cell growth measurements over different 3OC₆HSL input concentrations in a single experiment. The modified Lin's model implemented in this chapter can be written as

$$\mu(t) = \mu_{max} \left(\left(\frac{1}{1 + e^{-k_{in}(t-t_{in})}} \right) \left(\frac{1}{1 + e^{k_{de}(t-t_{de})}} \right) - e^{k_d(t-t_d)} \right), \quad (7.3)$$

$$\dot{x}_j(t) = \mu(t)x_j(t), \quad (7.4)$$

for $j = 1, \dots, M$ cell growth measurements corresponding to different constant input levels of 3OC₆HSL. The constants, k_d and t_d parameterises the decay dynamics in the cell growth.

"Single-cell" model

Generally, the CT-NARX model is obtained in a data-driven framework from regularly sampled input and output signals. In this investigation, before the alteration

explained above is imposed, the input and output signals to the systems are defined as $u(t) \in \mathbb{R}$ - 3OC₆HSL concentration and $y(t) \in \mathbb{R}$ - protein expression (GFP - BBa_T9002 system and RFP - "F2620-RC2" system) respectively. The structure of a general CT-NARX model can be defined by (Coca and Billings, 1999)

$$y^{n_i}(t) = f(\mathfrak{S}(t)) + e(t), \quad (7.5)$$

$$\mathfrak{S}(t) = (y(t), \dots, y^{n_i-1}(t), u(t), \dots, u^{n_i-1}(t)), \quad (7.6)$$

where n_i is the differential order, $f(\mathfrak{S}(t))$ is some unknown nonlinear function and $\mathfrak{S}(t) \in \mathbb{R}^{2n_i}$ is the model input vector of system input and output derivatives. The function $f(\cdot)$ can be described using a basis function decomposition

$$y^{n_i}(t) = \sum_{j=1}^{N_\theta} \theta_j \phi_j(\mathfrak{S}(t)), \quad (7.7)$$

where $\phi_j(\cdot)$ is a basis function with associated parameter, $\theta_j \in \mathbb{R}$. In this investigation, polynomial basis functions of maximum order $q = 3$ was used and second order system dynamics, $n_i = 2$ was assumed.

Alterations to the general form of the CT-NARX model was imposed, in order to accommodate a specialised form for this investigation by: (i) assigning the output signal to be the average ("single-cell") protein expression and (ii) only considering derivatives in the output signal and no cross-product terms between input and output signals.

This specialised form was implemented because: (i) the CT-NARX model is used to represent the dynamic function of the "single-cell" model that captures the deterministic part of the underlying "single-cell" protein expression dynamics of the systems and (ii) of the assumption that the input level of 3OC₆HSL was constant over the duration of each experiment, so the derivatives of the input signal were zero and the cross-product terms were unidentifiable (hence, in this investigation where constant input is assumed, higher order polynomial input transformations and cross-product terms between input and output signals are not used). As mentioned above there appeared to be a nonlinear gain variation in the "single-cell" protein expression responses associated with different input levels of 3OC₆HSL in an experiment, which is described using separate input gain terms k_j , for $j = 1, \dots, M$, resulting in the following modification of the CT-NARX model

$$\check{y}_j(t) = \frac{y_j(t)}{x_j(t)}, \quad (7.8)$$

$$\check{y}_j^{n_i}(t) = f(\check{y}(t), \dots, \check{y}^{n_i-1}(t)) + k_j u_j(t) + e(t), \quad (7.9)$$

for $j = 1, \dots, M$ "single-cell" protein expression signals corresponding to different constant input levels of 3OC₆HSL, where $\check{y}(t) \in \mathbb{R}$ is the average "single-cell" protein expression.

CT-NARX model with static input nonlinearity

The function $G(\cdot)$ was used to model the static nonlinear gain variation across input levels, which mapped the 3OC₆HSL input - $u(t)$ to the dynamic model input - $\tilde{u}(t)$, and the CT-NARX model was consequently modified to

$$\check{y}^{n_i}(t) = f(\check{y}(t), \dots, \check{y}^{n_i-1}(t)) + \tilde{u}(t) + e(t), \quad (7.10)$$

where $\tilde{u}(t) = G(u_*(t))$, $u_*(t) = \log_{10}(gu(t)) + 6$ (g is a scaling parameter which improves the numerical conditioning whilst prior estimation and model simulation, where $g = 1 \times 10^5$). Due to the log spacing in the levels of 3OC₆HSL, the log transformation was applied to the scaled input $gu(t)$ and the addition of 6 rescales it to a positive value. The function $G(\cdot)$ was described by the basis function decomposition

$$\tilde{u}(t) = \sum_{j=1}^B w_j \psi_j(u_*(t)), \quad (7.11)$$

where $w_j \in \mathbb{R}$ is the j^{th} basis function parameter, B is the number of basis functions, and in this investigation the radial basis functions were used, specifically the squared exponential function,

$$\psi_j(u_*(t)) = \exp\left(-\frac{1}{2\sigma_j^2} \|u_*(t) - \varphi_j\|_2^2\right), \quad (7.12)$$

where φ_j and σ_j are the respective centres and widths of the j^{th} basis function. Basis functions were centred on the levels of the input data values $u_*(t)$ and the corresponding width parameters were heuristically tuned in the range $\sigma_j \in [1, 1.5] \forall j$.

7.3.3 Parameter estimation and structure detection of system model

The identification of the system model of both BBa_T9002 and "F2620-RC2" systems, was undertaken by simultaneously implementing: (i) parameter estimation of the cell growth model and (ii) model estimation (parameter estimation and model structure detection (MSD)) of the "single-cell" model of each system. The parameter estimation and MSD was implemented using the two-stage model structure detection algorithm developed in Chapter 5 (Algorithm 5.3). The MSD algorithm applies a parameter significance test on the parameter distributions which are naturally generated as a byproduct of the approximate Bayesian computation (ABC) - sequential Monte Carlo (SMC) algorithm, followed by an information criterion test to obtain an improved model structure. The parameter estimation using ABC-SMC (Algorithm 5.1), is to iterate population estimates generated by basic ABC, gradually decreasing the error tolerance at each iteration. The posterior distribution at an iteration becomes the sampled prior distribution at the next iteration. Hence, the ABC-SMC algorithm reaches the target posterior in a sequential manner.

In implementing the two-stage MSD algorithm for the identification of the system models, the parameter size was set to $L = 200$ and the number of population iterations to $K = 3$ for the ABC-SMC parameter estimation step. The parameter distribution obtained provides the clear description of the uncertainty in model estimation given the limitation in the experimental data. It also includes the population heterogeneity observed due to variability in cell populations. The parameter vector obtained can be represented as

$$\boldsymbol{\theta} = (\mu_{max}, k_{in}, t_{in}, k_{de}, t_{de}, k_d, t_d, a_1, \dots, a_M, c_1, \dots, c_{N_\theta-1})^\top, \quad (7.13)$$

where $(\mu_{max}, k_{in}, t_{in}, k_{de}, t_{de}, k_d, t_d)$ are parameters of the modified Lin's model as shown in eqn(7.3 and 7.4); (a_1, \dots, a_M) are parameters associated with the dynamic function input to the "single-cell" model (static nonlinear gain) $\tilde{\mathbf{u}} = (a_1 u_1(t_0), \dots, a_M u_M(t_0))^\top$, where $u_j(t_0)$, for $j = 1, \dots, M$, corresponded to the rescaled input levels of 3OC₆HSL; and $(c_1, \dots, c_{N_\theta-1})$ are parameters associated with superset of model terms composed of polynomial transformations of $\check{y}(t)$ and its derivatives as shown in eqn(7.10). The MSD algorithm detected a parsimonious model structure composed of a reduced set of those terms in eqn(7.10). In this investigation the number of model terms in the superset (eqn(7.10)) was 9.

The model simulation step was performed by deterministic simulation of the mod-

els defined in eqn(7.3 and 7.4) - modified Lin's model and eqn(7.10) - CT-NARX model, using a fourth order Runge-Kutta method. The distance measure of simulations from the observations was obtained from the mean-sum-of-squared errors

$$d = \frac{1}{M} \sum_{i=1}^M \sum_{j=1}^{N_y} (\mathcal{Y}(j, i) - \mathcal{Y}^*(j, i))^2, \quad (7.14)$$

where $\mathcal{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_M]$ is the observed output matrix, $\mathbf{y}_j = (y_j(t_0), \dots, y_j(t_{N_y} - 1))^\top = (\check{y}_j(t_0)x_j(t_0), \dots, \check{y}_j(t_{N_y} - 1)x_j(t_{N_y} - 1))^\top$ for $j = 1, \dots, M$ and \mathcal{Y}^* is the simulated output matrix.

For the system model, the prior distribution of the parameters was defined as a uniform distribution. The prior of the "single-cell" model was scaled using least-squares (LS) parameter estimate obtained from the method of Coca and Billings (1999), whereas the prior of the cell growth model was scaled by graphical interpretation of the experimental data. The basis function parameters for the static nonlinear input function $G(\cdot)$ were estimated using LS from the target data $\tilde{\mathbf{u}} = (a_1 u_1(t_0), \dots, a_M u_M(t_0))^\top$, where $u_j(t_0)$, for $j = 1, \dots, M$, corresponded to the rescaled input levels of 3OC₆HSL.

Experiment 1 of both experimental datasets are used for identification purposes. In implementation, the experimental data (both cell growth and protein expression measurements) corresponding to the highest input level 3OC₆HSL = 1e-4 M for both systems, was omitted for identification purpose due to the data appearing to be a design outlier. The "single-cell" model was identified using the normalised "single-cell" protein expression responses. This was done in order to directly compare the static nonlinear input function $G(\cdot)$ of both BBa_T9002 and "F2620-RC2" systems.

7.4 Results and discussion

In using the collected experimental data of systems - BBa_T9002 and "F2620-RC2", the main objective of this chapter is to determine if the functional module - BBa_F2620 is robustly characterised, regardless of not being able to directly monitor its dynamics. In undertaking this investigation, the identification of system models for both systems was achieved, which constituted of a "single-cell" model and a cell growth model. The computational Bayesian identification framework was chosen to be implemented here, instead of the data-driven framework used in Chapter 4 because the former is capable of quantifying and explaining the vari-

ability phenomenon observed in different cell populations robustly, that could aid design procedures. The "single-cell" models capture the underlying "single-cell" protein expression dynamics of the systems, which are transformed into frequency functions for spectral analysis, that characterises the systems under study. Additionally, the cell growth model is used to provide a solution to explicitly quantify the cellular burden caused by each system and the parameter distribution naturally generated by the identification framework is used to characterise the population heterogeneity observed.

The identification of the system models for both BBa_T9002 and "F2620-RC2" systems, was undertaken by simultaneously implementing: (i) parameter estimation of the cell growth model and (ii) model estimation (parameter estimation and model structure detection) of the "single-cell" model using the two-stage model structure detection algorithm developed in Chapter 5 (Algorithm 5.3). The cell growth model predicts the growth responses corresponding to different input levels of 3OC₆HSL in a single experiment. The number of parameters governing the cell growth model is 7 (eqn(7.3 and 7.4)). The MSD of the "single-cell" model was implemented to detect a parsimonious structure to represent the average protein expression of a single cell. The number of possible model terms used for representing the "single-cell" model is 9 (eqn(7.10)). The "population-level" protein expressions of the systems, which are the output signals to the system models and used in the distance measure of the ABC-SMC step, are obtained by multiplying the average "single-cell" protein expressions by their corresponding growth responses. Experiment 1 of both experimental datasets are used for identification purposes. In implementation, the experimental data (both cell growth and protein expression measurements) corresponding to the highest input level 3OC₆HSL = 1e-4 M for both systems, was excluded in the identification step due to it appearing to be an outlier. Hence, the number of parameters associated with the static nonlinear gain (input signal to the dynamic function of the "single-cell" model) was 7 (eqn 7.11).

7.4.1 Cell growth properties

The maximum values the growth responses of BBa_T9002 system, when induced by different input levels of 3OC₆HSL varies between 0.45 - 0.6, which are initially attained in approximately 6 hours (Figure 6.5A and 7.3A). When higher GFP expression is observed, as the BBa_T9002 system is induced by higher concentration of 3OC₆HSL, a lower growth response is consequently observed. This is due to cells prioritising cellular resources for protein expression more than cell divi-

Table 7.2: Parameters (mean values) of the cell-growth models for both BBa_T9002 and "F2620-RC2" systems.

Parameters	BBa_T9002	"F2620-RC2"
μ_{max}	0.05	0.02
k_{in}	0.12	0.03
t_{in}	0.02	0.02
k_{de}	26.90	4.11
t_{de}	38.91	149.54
k_d	0.01	0.01
t_d	983.59	1.21×10^3

sion. The maximum values the growth responses of "F2620-RC2" system, when induced by different input levels of 3OC₆HSL varies between 0.6 - 1.2, which are initially attained in approximately 13 hours (Figure 6.6A and 7.4A). However, when higher RFP expression is observed, as the "F2620-RC2" system is induced by higher concentration of 3OC₆HSL, a lower growth response is not always consequently observed. The reason for this unusual behaviour cannot be explained in this investigation. The following is suggested, the difference between the two systems are their respective reporter cascades and the fluorescence proteins which make up these reporter cascades, GFP and RFP, have very similar properties (Table 7.1). Therefore the expression of these proteins should not have differing effect on the cellular resources of the cells. However, the ribosome binding sites (RBSs) of the respective reporter cascades have varying relative strength, which influences the amount of proteins expressed. The RBS acts as a control mechanism on the amount of protein expressed, thereby influencing the cellular resources used in protein expression and indirectly affecting cell division.

In summary, the BBa_T9002 system has a lower maximum value of growth response in comparison to the "F2620-RC2" system when induced by the same concentration of 3OC₆HSL, however, the BBa_T9002 system attains its maximum value of growth response in a much quicker time than the "F2620-RC2" system. This is an important feature of the systems, as this effects the "population-level" protein expression of the systems respectively. In this investigation, the parameters of the cell growth models are used to explicitly quantify the usage of cellular resources by the systems, which is influenced by their respective RBSs. The mean parameter values of the cell growth models are shown in Table 7.2.

Table 7.3: Parameters (mean values) of the "single-cell" models (CT-NARX) for both BBa_T9002 and "F2620-RC2" systems.

Parameters	BBa_T9002	"F2620-RC2"
c_1	-1.47×10^{-4}	-2.64×10^{-5}
c_2	-0.03	—
c_3	-0.13	-0.04
c_4	-0.18	-0.10
c_5	30.84	—

7.4.2 "Single-cell" protein expression properties

The "single-cell" models (CT-NARX) of BBa_T9002 and "F2620-RC2" systems identified using the two-stage MSD algorithm (Algorithm 5.3) are

$$\begin{aligned} \ddot{y}(t) &= c_1\dot{y}(t) + c_2\ddot{y}(t) + c_3\dot{y}(t)\dot{y}(t) + c_4\dot{y}^2(t)\dot{y}(t) + c_5\dot{y}^3(t) + \ddot{u}(t) \text{ and} \\ \ddot{y}(t) &= c_1\dot{y}(t) + c_3\dot{y}(t)\dot{y}(t) + c_4\dot{y}^2(t)\dot{y}(t) + \ddot{u}(t) \text{ respectively,} \end{aligned}$$

where $\dot{y}(t)$ was the model output signal - average "single-cell" protein expression, with associated parameters c_1, c_2, c_3, c_4 and c_5 . The input term $\ddot{u}(t)$ was obtained from a static transformation $G(\cdot)$ of the input signal 3OC₆HSL, which was primarily used to describe the static switching effect in dynamics across linearly increasing levels of 3OC₆HSL (see above). The "single-cell" models were identified using the normalised "single-cell" protein expression responses. This was done in order to directly compare the static nonlinear input functions $G(\cdot)$ of both BBa_T9002 and "F2620-RC2" systems. The mean values of the parameters of the "single-cell" models for BBa_T9002 and "F2620-RC2" systems are summarised in Table 7.3.

Nonlinear static gain properties

The static nonlinear input functions $G(\cdot)$ of both BBa_T9002 and "F2620-RC2" systems can be directly compared from Figure 7.3B and 7.4B. The extra damping terms $\dot{y}(t)$ and $\dot{y}^3(t)$ present in the "single-cell" model of the BBa_T9002 system, compared to the "F2620-RC2" system, helps to capture the BBa_T9002 system's faster transient responses and shorter times in achieving steady state, in both "single-cell" and "population-level" protein expression responses corresponding to different input levels of 3OC₆HSL (Figure 7.3C,D and 7.4C,D). This could be the reason for higher nonlinear static gains in the BBa_T9002 system, as the identification framework is data-driven, higher gains are required to achieve faster transient responses given a well damped system.

The spread and variation on the nonlinear static gains corresponding to linearly increasing levels of 3OC₆HSL, for the BBa_T9002 system is wider and spaced out in comparison to that of the "F2620-RC2" system. This is clearly reflected in Figure 7.3C and D, as the protein expression responses to increasing input levels of 3OC₆HSL, are distinguishable and gradually increasing. This is not observed for the "F2620-RC2" system (Figure 7.4C and D), as there is a distinctive bimodality separation amongst the protein expression responses to increasing input levels of 3OC₆HSL. This could be due to the prioritised demand imposed on the cellular resources by the RBS of the "F2620-RC2" system, for the expression of RFP. Equally, it could be due to the consequence of using less sensitive optical filters for measuring the RFP expression.

Dynamic function properties

As mentioned above, the extra terms $\dot{y}(t)$ and $\dot{y}^3(t)$ are present in the "single-cell" model of the BBa_T9002 system, compared to the "F2620-RC2" system. The BBa_T9002 system has faster transient responses and shorter times in achieving steady state, in both "single-cell" and "population-level" protein expression responses corresponding to different input levels of 3OC₆HSL (Figure 7.3C,D and 7.4C,D). In the time domain, the BBa_T9002 system's "single-cell" protein expression dynamics is not comparably similar to the "single-cell" protein expression dynamics of the "F2620-RC2" system by visualisation.

The analysis of nonlinear system in the frequency domain can provide important insights into a system's nonlinearity and physical behaviour. Capturing the system dynamics using only the input and output signal, to study the relationship between them using spectral analysis is important, which was achieved by deriving the "single-cell" (CT-NARX) model. Generalised frequency response functions (GFRFs) are higher-order functions that are multi-dimensional and used in representing nonlinear systems in the frequency domain. The probing method is used to compute the GFRFs of the parametric model - "single-cell" CT-NARX model by excluding the noise term. The computation of the first-order and second-order GFRFs of the "single-cell" (CT-NARX) models of both BBa_T9002 and "F2620-RC2" systems is given in Appendix A and Appendix B respectively.

The bode plots (magnitude and phase) of the first-order GFRFs of both BBa_T9002 and "F2620-RC2" systems are shown in Figure 7.5. The time-domain models are identified using experimental data with time interval of one data sample.

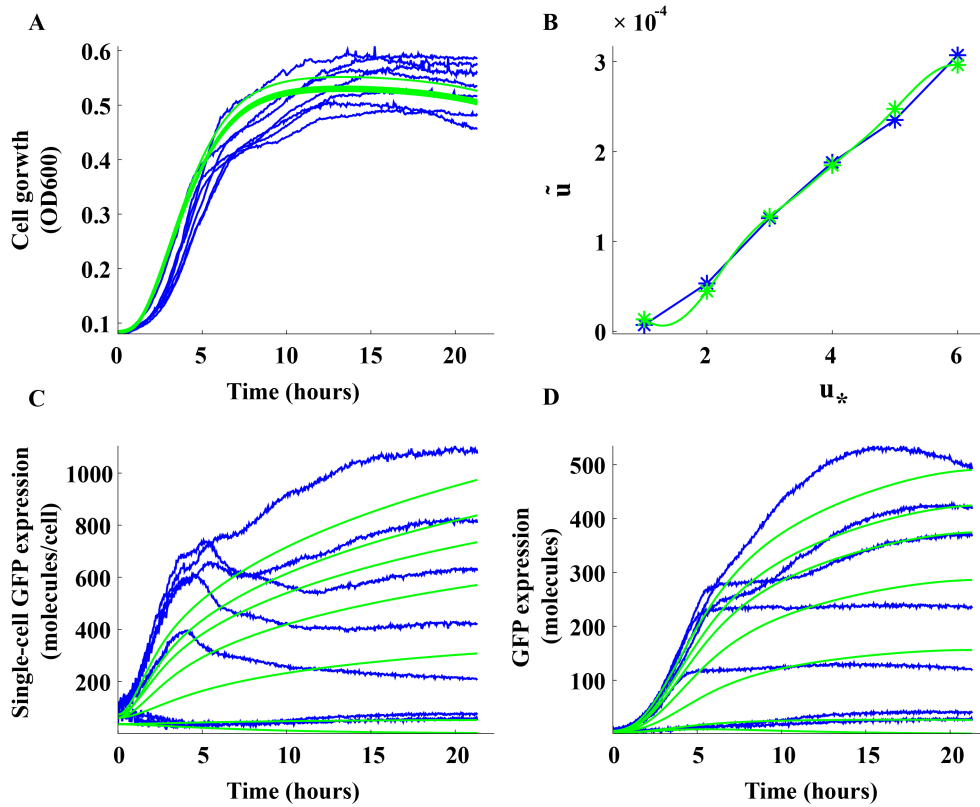


Figure 7.3: Mean model simulation for BBa_T9002 system. A. Comparison of growth response (blue) and the modified Lin's model prediction (green), B. Static model of the input nonlinearity, estimate of the static function $G(\cdot)$ (green) compared to the experimental data curves (blue), C. Comparison of "single-cell" GFP expression (blue) and "single-cell" model prediction (green) and D. Comparison of GFP expression (blue) and system model prediction (green). Note that the response corresponding to the input levels $3OC_6HSL = 0$ (in B) and $1e-4 M$ (all) have been omitted because of the log transformation and outlier observation respectively.

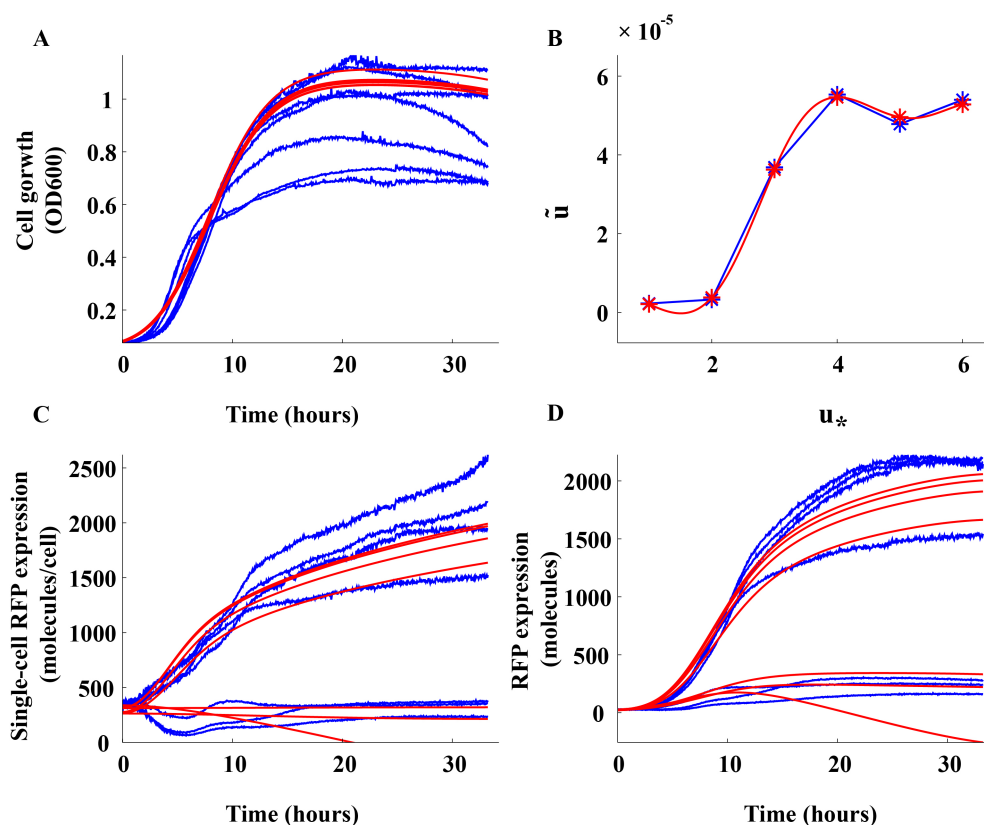


Figure 7.4: Mean model simulation for "F2620-RC2" system. A. Comparison of growth response (blue) and the modified Lin's model prediction (red), B. Static model of the input nonlinearity, estimate of the static function $G(\cdot)$ (red) compared to the experimental data curves (blue), C. Comparison of "single-cell" RFP expression (blue) and "single-cell" model prediction (red) and D. Comparison of RFP expression (blue) and system model prediction (red). Note that the response corresponding to the input levels $3OC_6HSL = 0$ (in B) and $1e-4 M$ (all) have been omitted because of the log transformation and outlier observation respectively.

Therefore it was necessary to rescale the frequency axis (x -axis) from (rad/sec) to (rad/sec) $\times \frac{1}{T}$. The magnitude plots of the first-order GFRFs of both BBa_T9002 and "F2620-RC2" systems have unique properties of their own. There is a resonance spike noticeable in the "F2620-RC2" system's magnitude plot, which the data-driven identification framework would have used to capture the oscillatory-type "single-cell" protein expression as seen in Figure 7.4C, whereas the dynamics of the BBa_T9002 system is more like a first-order step response at the "single-cell" level (Figure 7.3C). The second-order GFRFs of both systems have their first-order GFRFs repeated across the second frequency axes forming three dimensional plots (slicing across the second frequency axis provides the two dimensional first-order GFRF of the system). The simulated prediction of the "single-cell" protein expression of the "F2620-RC2" system does not oscillate even though its first-order GFRF entails a resonance, as the resultant output frequency response (which involves $H_1(\omega_1)$, $H_2(\omega_1, \omega_2)$ and $H_3(\omega_1, \omega_2, \omega_3)$) is damped because of the interference of $H_2(\omega_1, \omega_2)$ and $H_3(\omega_1, \omega_2, \omega_3)$ on the first-order GFRF. This is the attractive property of the spectral analysis using GFRFs, the effect on the dynamics of the systems caused by the model terms and parameters can be interpreted, which aids control design of systems in control engineering. The full interrogation on how the different frequency orders combine to produce the resulting time-domain response is not undertaken here (future work), however, the usefulness of the GFRFs for design purposes is shown thereby permitting another tool for dynamic characterisation of genetic parts.

From the observation seen in the collected experimental data, the BBa_T9002 system has faster transient responses and shorter times in achieving steady state in "single-cell" protein expression responses corresponding to different input levels of 3OC₆HSL, than the "F2620-RC2" system. In the time-domain, the model structures and parameters of both systems are not very different and can be represented similarly (taking into account that the parameters c_2 and c_5 are zero for "F2620-RC2" system). This unique representation whose model terms along with their respective parameters combine nonlinearly to produce the different time responses of BBa_T9002 and "F2620-RC2" systems (the delay). This indicates that, when the nonlinear effects introduced by the nonlinear static gain is removed, the inherent dynamic characteristics of both systems have commonality. This proves to be a strong and promising result, which shows that the reporter cascades do influence the "relative" dynamics of the systems and characterising only the functional module - BBa_F2620 relative to a reporter cascade as an unachievable task using the implemented investigation in this thesis. However, with the identifi-

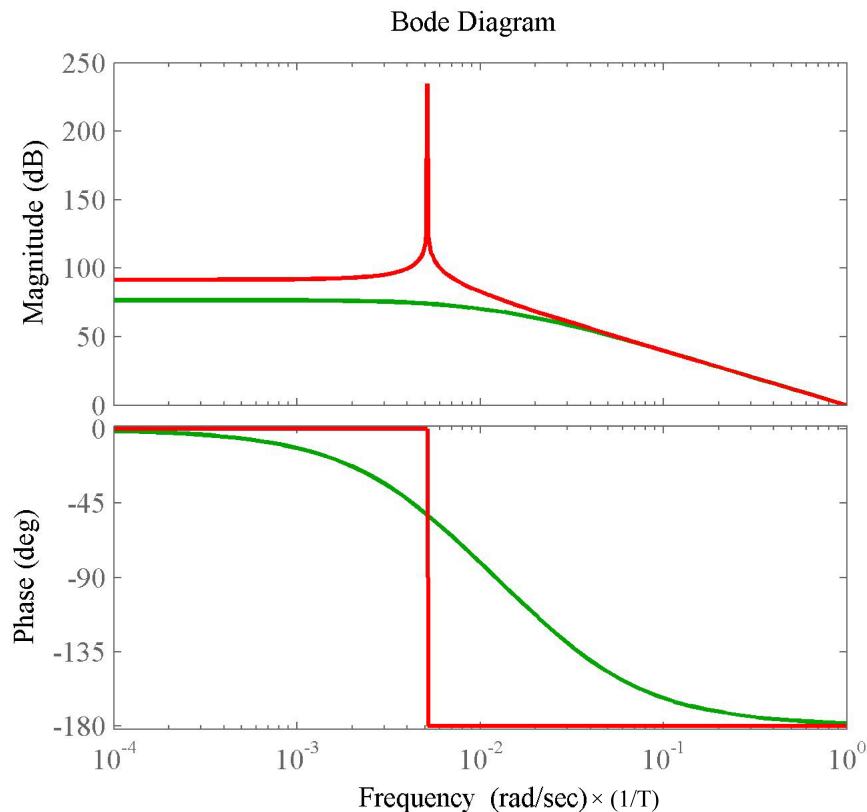


Figure 7.5: Bode plots (both magnitude and phase) of the first-order GFRFs of both BBa_T9002 (green) and "F2620-RC2" (red) systems, using mean values of the parameters.

cation and analysis tools used, the commonality of the systems (same functional module linked to two different reporter cascades) is retrieved and adequately characterised.

As stated earlier, the RBSs of the reporter cascades act as a control mechanism on the amount of protein expressed, thereby influencing the cellular resources used in protein expression and indirectly affecting cell division. Therefore, the reporter cascade's influence on the "relative" dynamics of the systems are mostly due to the RBSs, which in effect is causing the delay observed in the system's dynamics and the varying nonlinear switching behaviour. The RBS of the "F2620-RC2" system prioritises demand on the cellular resources for protein expression more than the RBS of the BBa_T9002 system, indirectly causing a delay in the growth of the "F2620-RC2" system. However, the maximum growth of the "F2620-RC2" system is greater than that of the BBa_T9002 system. This is due to less protein being produced by the "F2620-RC2" system, which could be caused by RNA stability or

codon optimisation (Goodman et al., 2013).

As demonstrated in Ellis et al. (2009), where static analysis is used to quantify the relative strength of promoters, here dynamic analysis can be used to quantify the relative strength of RBSs, given the suggestions mentioned in this chapter holds true: (i) the fluorescence proteins used here do not affect the "relative" dynamics of the systems and (ii) the unique representation of the "single-cell" protein expression dynamics in the time-domain captures the "single-cell" protein expression of a system made up of the functional module - BBa_F2620 and an arbitrary reporter cascade. This could help to predict dynamic properties the in model-based design. There is a case for these dynamic properties to be added as an extension to the datasheets of the respective genetic parts. The two crucial properties needed to be reported are: (i) the nonlinear static behaviour and (ii) the "single-cell" time-domain and frequency-domain models along with the parameters c_1 , c_2 , c_3 , c_4 and c_5 . This can only be validated by conducting more experiments with systems made of the functional module - BBa_F2620 and different reporter cascades (future work). Until then, the delay effect introduced by the RBS cannot be conclusively assured, the commonality of systems made up of the functional module - BBa_F2620 and different arbitrary reporter cascades is not guaranteed, thereby leaving the debate - appropriateness of reporter cascades for the use of characterisation of genetic parts, open to the synthetic biology community.

Also, in this investigation the dynamics of the functional module is not decoupled from the "relative" dynamics of the system, the reported properties may be expected to hold true when the functional module is chosen for building a higher-order system. Relative calibration using new experimental data of the higher-order system as shown in Ellis et al. (2009) will be required though (discussed in the introduction).

7.4.3 Model validation of a unified model with parameter uncertainty

The mean model simulation of both systems are shown in Figure 7.3 and 7.4, which includes the growth, "single-cell" protein expression and "population-level" protein expression responses. This was achieved by simulating the system models using the mean values of the estimated parameter distributions (eqn 7.13) which are shown in Table 7.2 and 7.3 (the mean values of the static nonlinear gains can be seen in Figure 7.3B and 7.4B). Experiment 1 of both experimental datasets was used for identification purposes. However, the remaining 8 experiments in each experimental dataset was reserved for validation purposes. The identified

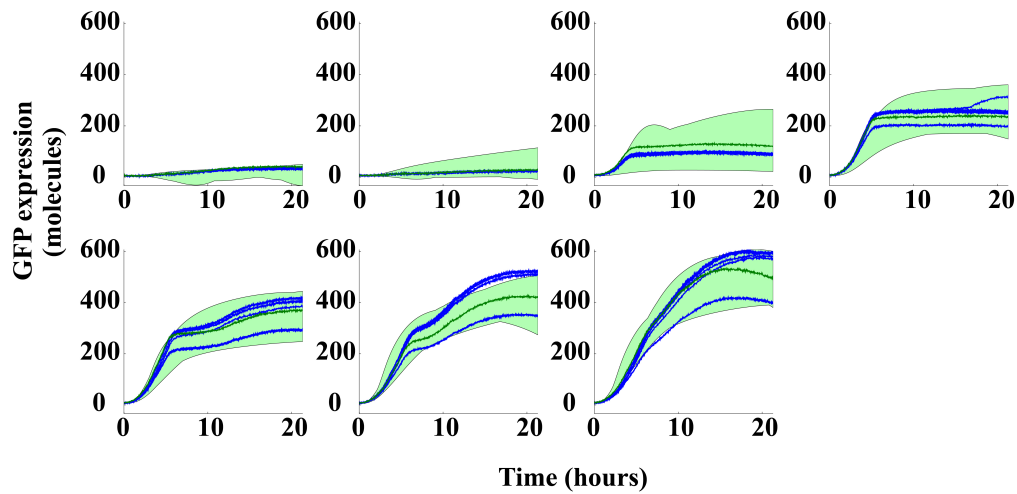


Figure 7.6: BBa_T9002 system model variation. Comparison of GFP expression used for identification purpose (dark green line), GFP expression of other experiments (blue) and identified system model (light green shaded region - indicates uncertainty from the ABC parameter range). The plots corresponds to responses to increasing input level from left to right - top to bottom.

system model of each system, which was achieved using only one experiment, is simulated using all the parameter samples obtained from the distributions that were naturally generated by the identification framework. The simulations are shown in Figure 7.6 and 7.7. One can observe that most experimental data falls between the shaded region of uncertainty implying confidence in the identified model and the estimated parameter distribution. This also helps to capture the population heterogeneity of the cell populations. The variation observed in the time domain as shown, can be translated into the frequency domain as well to aid design, which is hugely advantageous, as the computation of the GFRFs are done from a parametric representation.

7.5 Summary

The novel experimentation in this chapter is carried out to investigate if reporter cascades are appropriate for characterisation. The reporter cascades are shown to have influential effect on the "relative" dynamics of both systems under investigation, which is suggested to be caused by the RBSs. Characterising only the functional module - BBa_F2620 relative to a reporter cascade was proven to be an unachievable task using the implemented investigation in this thesis. In addition, the identification and analysis tools used in this chapter, was capable of retrieving and adequately characterising the invariant features common to both systems

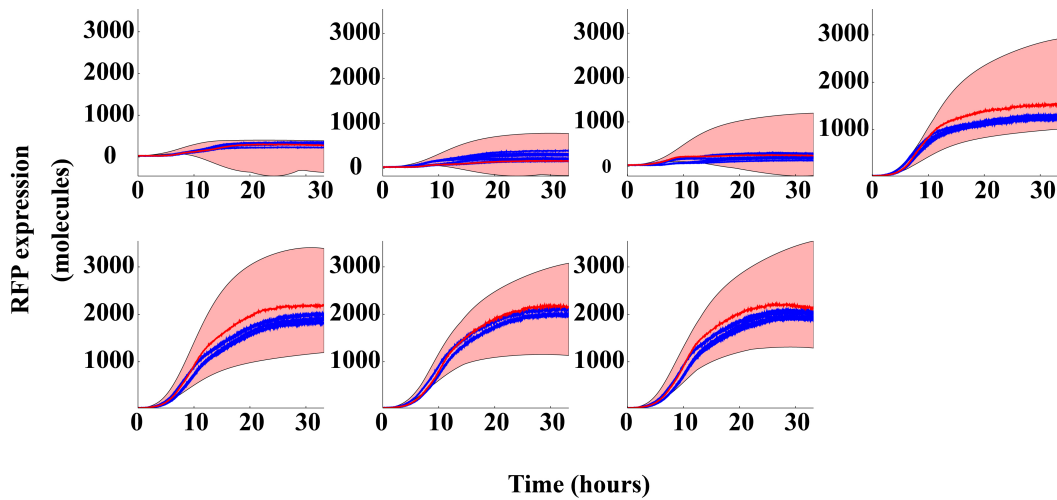


Figure 7.7: "F2620-RC2" system model variation. Comparison of RFP expression used for identification purpose (dark red line), RFP expression of other experiments (blue) and identified system model (light red shaded region - indicates uncertainty from the ABC parameter range). The plots corresponds to responses to increasing input level from left to right - top to bottom.

under investigation (same functional module linked to two different reporter cascades). It was also concluded that, more experiments with systems made of the functional module - BBa_F2620 and different reporter cascades have to be conducted, to deduce if the unique representation in time-domain of the "single-cell" protein expression dynamics obtained in this chapter can be used to represent a system made of the functional module - BBa_F2620 and an arbitrary reporter cascade to aid model-based design. Therefore, leaving the debate - appropriateness of reporter cascades for the use of characterisation of genetic parts, open to the synthetic biology community. These results hint at the possibility that dynamic characterisation with predictive ability can lead to new design tools in synthesising functional bioparts and devices.

Additional dynamic characterisation is also demonstrated such as the estimation of parameter uncertainty to capture cell population heterogeneity and explicit cellular burden quantification using a cell growth model. Therefore, to design a biological system efficiently, the following are needed to assist model-based design: (i) robust characterisation of the genetic part, (ii) quantification of the expected variability by the cell population and (iii) the prediction of system's microbial growth. The work presented in this chapter has shown the approach in doing this and, the relevant identification and analysis tools needed to achieve this design goal.

Chapter 8

Conclusions

8.1 Conclusions and summary

Overall, this thesis introduced a new methodology for characterisation and analysis of the dynamics of genetic parts - using newly designed experimentation, nonlinear system identification and frequency domain analysis. The models and analysis presented can provide an extension to the datasheets of the respective genetic parts to aid model-based design.

This thesis has proposed a nonlinear dynamic modelling framework and developed another to be used, to achieve robust dynamic characterisation of genetic parts in biological systems. These frameworks are popular in the field of control engineering but are novel to the field of synthetic biology. The dynamic characterisation is demonstrated at the "population-level", as population of cells are used to complete the required and desired tasks, even though a cell is engineered to acquire a certain desired functionality.

The investigation, whether a reporter cascade is appropriate for the characterisation of genetic parts is also undertaken. This is important to the synthetic biology community because functional modules can rarely be directly monitored and they are characterised with respect to reporter cascades used for monitoring the system dynamics. However, when higher-order genetic parts are designed, different functional modules are usually synthesised and ligated together, while excluding the reporter cascades used during characterisation.

In Chapter 4, a data-driven modelling framework that utilises a regression approach is proposed to identify nonlinear black-box models for dynamic character-

isation of genetic parts in biological systems. The framework was demonstrated on a transcriptional regulatory genetic part - BB_T9002, for which a "population-level" continuous-time (CT) nonlinear autoregressive model with exogenous input (NARX) was obtained, that was compact and had accurate predictions of the experimental data. The CT-NARX model was also benchmarked against dynamic and static biochemical models, which were based on an enzymatic reaction scheme. The enzymatic reaction scheme model was shown to be inaccurate and inconsistent with its associated simplified form - the Hill equation. A transcriptional regulatory system was used as a case study because it is one of the most simplistic genetic functional module and which is frequently used as the foundational module to design higher-order genetic parts. The following was clearly established in Chapter 4, the need to: (i) gather more experimental data of the BB_T9002 system and (ii) develop a principled approach to quantify the cell population variability observed in the experimental data. These were required, in order to robustly characterise a transcriptional regulatory system.

A computational Bayesian identification framework for nonlinear CT systems that utilises a simulation approach as opposed to a regression approach is developed. The main contribution of this algorithm to the suite of methods available for CT nonlinear system identification is the signal derivative free approach and the estimation of the model parameter uncertainty by constructing a distribution. The identification algorithm uses the approximate Bayesian computation (ABC) sequential Monte Carlo (SMC) method, which is a rejection sampling technique for inferring parameters of a model. Parameter distributions intrinsically generated by ABC-SMC estimation algorithm is used to drive term selection by significance testing. The simulation results shown in Chapter 5 demonstrate the high fidelity of the ABC approach to increase in noise levels in the measurements. The developed identification framework will aid the quantification and characterisation of variability in gene expression that is observed in different cell populations, in a principled way.

Further experimentation is demonstrated in Chapter 6 were: (i) a novel investigation is implemented by assembling a new genetic part "F2620-RC2", which will help to investigate if a reporter cascade has an influential effect on the "relative" dynamics of the system it has been linked to and (ii) cell growth and protein expression measurements for longer time period (lag phase to death phase) are collected for systems - BB_T9002 and "F2620-RC2", in order to capture and model the full range of dynamics exhibited by transcriptional regulatory systems. The

BB_T9002 and "F2620-RC2" systems share the same functional module - BB_F2620, however, they have different reporter cascades - BB_E0240 and BB_J06702 respectively.

The collected experimental data were used to derive dynamic system models of both BB_T9002 and "F2620-RC2" systems. Each system model, consisted of: (i) a cell growth model, which captures the growth responses of the respective system over 8 different 3OC₆HSL input concentrations and (ii) a "single-cell" model, which describes the average number of fluorescence proteins expressed in each cell when induced by different 3OC₆HSL input concentrations. The parameters of the cell growth models were used to explicitly quantify the usage of cellular resources by the systems under investigation. The identified "single-cell" models were transformed into frequency models by computing their generalised frequency response functions (GFRFs), which serves as an alternative tool for dynamic characterisation of the genetic parts for design purposes. In the time-domain, the model structures and parameters of both systems are not very different and can be represented similarly. This unique representation whose model terms along with their respective parameters combine nonlinearly to produce the different time responses of BBa_T9002 and "F2620-RC2" systems. This behaviour is suggested to be caused by the ribosome binding site (RBS), which imposes a control mechanism on the protein expression of the system. However, when the nonlinear effects introduced by the nonlinear static gain is removed, the inherent dynamic characteristics of both systems have commonality. This indicates that the reporter cascades do influence the "relative" dynamics of the systems and characterising only the functional module - BBa_F2620 relative to a reporter cascade as an unachievable task using the implemented investigation in this thesis. However, with the identification and analysis tools used, the commonality of the systems (same functional module linked to two different reporter cascades) is retrieved and adequately characterised. More experiments with systems made of the functional module - BBa_F2620 and different reporter cascades have to be conducted, to deduce if the unique representation in time-domain of the "single-cell" protein expression dynamics obtained in Chapter 7 can be used to represent a system made of the functional module - BBa_F2620 and an arbitrary reporter cascade to aid model-based design. Therefore, leaving the debate - appropriateness of reporter cascades for the use of characterisation of genetic parts, open to the synthetic biology community.

8.2 Future work

This thesis has made an important contribution towards the dynamic characterisation of transcriptional regulatory genetic parts. It also provides a starting point for further work in the field.

- In this thesis, predictive model-based design is emphasised to overcome challenges in designing higher-order genetic parts. Two transcriptional regulatory systems are designed, to demonstrate the effect of reporter cascades on the "relative" dynamics of the systems. It was concluded that the influential effect observed in the "relative" dynamics is due to the RBSs. The time-domain and frequency-domain models which capture the key properties of the dynamics were presented, that showcases the effect caused by the RBS. In order to validate this study, additional systems should be built with the same functional module but different reporter cascades whose RBS strength varies. The time-domain and frequency-domain models presented here should be used to predict their dynamics and be validated against experimental data. This could provide quantification of the relative strengths of RBSs as discussed in Chapter 7, which will aid model-based design.
- The transcriptional regulatory genetic parts do have simpler dynamics compared to some other existing genetic parts, such as the repressilator (Elowitz and Leibler, 2000). In order to robustly characterise genetic parts with high dynamics, the systems are required to be persistently excited, which is presently unachievable using chemical inputs. However, the recent advancement in optogenetics allow different wavelengths of light to externally excite the biological systems consisting of light-active proteins. Additional advantages of using optogenetics are light travels faster than small molecules which takes time to diffuse and light can also be directed at specific parts of a cell (Baker, 2012).
- The computational Bayesian identification framework presented in Chapter 5 does model structure detection (MSD) and model selection (MS) based on derivative order separately. These two algorithms could be merged to produce one integrated algorithm which solves both challenges. This could be achieved by assigning weights to both models and parameters as shown in Toni et al. (2009). The Cha-Srihari measure that is used to detect which model parameter distributions have evolved the most from their priors, could be incorporated into the ABC-SMC estimation step to provide an importance sampling step of the parameters. This could be used to improve the MSD

procedure computationally. The problem relating to nonlinear system identification is the large dimension of terms in the model space which has been highlighted by Ninness (2009). The ABC-SMC needs to run multiple times to get a final posterior distribution of the model parameters. This implies a high computational burden, that can be solved using parallelisation, as done in this thesis using multi-core processing. However, even greater speed can be achieved by using new graphics based technologies.

Building on the foundations laid in this thesis new design tools on synthesising functional bioparts and devices can be developed in the future.

Appendix A

GFRFs computation of the "single-cell" model of BBa_T9002 system

The "single-cell" model (CT-NARX) of BBa_T9002 system identified using the two-stage MSD algorithm was

$$\ddot{y}(t) = c_1\dot{y}(t) + c_2\ddot{y}(t) + c_3\dot{y}(t)\dot{y}(t) + c_4\dot{y}^2(t)\dot{y}(t) + c_5\dot{y}^3(t) + \tilde{u}(t) \quad (\text{A.1})$$

First order, $N_f = 1$

$$\tilde{u}(t) = \sum_{i=1}^{N_f} e^{j2\pi f_i t} = e^{j2\pi f_1 t}. \quad (\text{A.2})$$

$$\dot{y}(t) = \sum_{i=1}^{N_f} \dot{y}_i(t) = \dot{y}_1(t). \quad (\text{A.3})$$

$$\dot{y}_i(t) = \sum_{i_1=1}^{N_f} \dots \sum_{i_{N_f}=1}^{N_f} H_i(f_{i_1}, \dots, f_{i_{N_f}}) e^{j2\pi(f_{i_1} + \dots + f_{i_{N_f}})t}, \quad (\text{A.4})$$

$$\dot{y}_1(t) = H_1(f_1) e^{j2\pi f_1 t}. \quad (\text{A.5})$$

$$\therefore \dot{y}(t) = H_1(f_1) e^{j2\pi f_1 t}, \quad (\text{A.6})$$

$$\ddot{y}(t) = j2\pi f_1 H_1(f_1) e^{j2\pi f_1 t}, \quad (\text{A.7})$$

$$\ddot{y}(t) = (j2\pi f_1)^2 H_1(f_1) e^{j2\pi f_1 t}. \quad (\text{A.8})$$

$$(\text{A.9})$$

By substituting the frequency transforms into the model (eqn(A.1)) and equating to $e^{j2\pi f_1 t}$ provides the first-order function,

$$H_1(\omega_1) = \frac{1}{(j\omega_1)^2 - c_2 j\omega_1 - c_1}. \quad (\text{A.10})$$

where $\omega_1 = 2\pi f_1$.

Second order, $N_f = 2$

$$\tilde{u}(t) = \sum_{i=1}^{N_f} e^{j2\pi f_i t} = e^{j2\pi f_1 t} + e^{j2\pi f_2 t}. \quad (\text{A.11})$$

$$\check{y}(t) = \sum_{i=1}^{N_f} \check{y}_i(t) = \check{y}_1(t) + \check{y}_2(t). \quad (\text{A.12})$$

$$\check{y}_i(t) = \sum_{i_1=1}^{N_f} \dots \sum_{i_{N_f}=1}^{N_f} H_i(f_{i_1}, \dots, f_{i_{N_f}}) e^{j2\pi(f_{i_1} + \dots + f_{i_{N_f}})t}, \quad (\text{A.13})$$

$$\check{y}_1(t) = H_1(f_1)e^{j2\pi f_1 t} + H_1(f_2)e^{j2\pi f_2 t}, \quad (\text{A.14})$$

$$\check{y}_2(t) = H_2(f_1, f_1)e^{j2\pi(2f_1)t} + 2H_2(f_1, f_2)e^{j2\pi(f_1+f_2)t} + H_2(f_2, f_2)e^{j2\pi(2f_2)t}. \quad (\text{A.15})$$

$$\begin{aligned} \therefore \check{y}(t) &= H_1(f_1)e^{j2\pi f_1 t} + H_1(f_2)e^{j2\pi f_2 t} + H_2(f_1, f_1)e^{j2\pi(2f_1)t} \\ &\quad + 2H_2(f_1, f_2)e^{j2\pi(f_1+f_2)t} + H_2(f_2, f_2)e^{j2\pi(2f_2)t}, \end{aligned} \quad (\text{A.16})$$

$$\dot{\check{y}}(t) = j2\pi(f_1 + f_2)2H_2(f_1, f_2)e^{j2\pi(f_1+f_2)t} \text{ only required}, \quad (\text{A.17})$$

$$\ddot{\check{y}}(t) = \left(j2\pi(f_1 + f_2)\right)^2 2H_2(f_1, f_2)e^{j2\pi(f_1+f_2)t} \text{ only required}. \quad (\text{A.18})$$

By substituting the frequency transforms into the model (eqn(A.1)) and equating to $2e^{j2\pi(f_1+f_2)t}$ provides the second-order function,

$$H_2(\omega_1, \omega_2) = \frac{0.5c_3 \left(H_1(\omega_1)H_1(\omega_2)j\omega_2 + H_1(\omega_1)H_1(\omega_2)j\omega_1 \right)}{\left(j(\omega_1 + \omega_2) \right)^2 - c_2 \left(j(\omega_1 + \omega_2) \right) - c_1}. \quad (\text{A.19})$$

Appendix B

GFRFs computation of the "single-cell" model of "F2620-RC2" system

The "single-cell" model (CT-NARX) of "F2620-RC2" system identified using the two-stage MSD algorithm was

$$\ddot{y}(t) = c_1 \dot{y}(t) + c_3 \dot{y}(t) \dot{y}(t) + c_4 \dot{y}^2(t) \dot{y}(t) + \tilde{u}(t) \quad (\text{B.1})$$

First order, $N_f = 1$

$$\tilde{u}(t) = \sum_{i=1}^{N_f} e^{j2\pi f_i t} = e^{j2\pi f_1 t}. \quad (\text{B.2})$$

$$\dot{y}(t) = \sum_{i=1}^{N_f} \dot{y}_i(t) = \dot{y}_1(t). \quad (\text{B.3})$$

$$\dot{y}_i(t) = \sum_{i_1=1}^{N_f} \dots \sum_{i_{N_f}=1}^{N_f} H_i(f_{i_1}, \dots, f_{i_{N_f}}) e^{j2\pi(f_{i_1} + \dots + f_{i_{N_f}})t}, \quad (\text{B.4})$$

$$\dot{y}_1(t) = H_1(f_1) e^{j2\pi f_1 t}. \quad (\text{B.5})$$

$$\therefore \dot{y}(t) = H_1(f_1) e^{j2\pi f_1 t}, \quad (\text{B.6})$$

$$\dot{\dot{y}}(t) = j2\pi f_1 H_1(f_1) e^{j2\pi f_1 t}, \quad (\text{B.7})$$

$$\ddot{y}(t) = (j2\pi f_1)^2 H_1(f_1) e^{j2\pi f_1 t}. \quad (\text{B.8})$$

$$(\text{B.9})$$

By substituting the frequency transforms into the model (eqn(B.1)) and equating to $e^{j2\pi f_1 t}$ provides the first-order function,

$$H_1(\omega_1) = \frac{1}{(j\omega_1)^2 - c_1}. \quad (\text{B.10})$$

where $\omega_1 = 2\pi f_1$.

Second order, $N_f = 2$

$$\tilde{u}(t) = \sum_{i=1}^{N_f} e^{j2\pi f_i t} = e^{j2\pi f_1 t} + e^{j2\pi f_2 t}. \quad (\text{B.11})$$

$$\check{y}(t) = \sum_{i=1}^{N_f} \check{y}_i(t) = \check{y}_1(t) + \check{y}_2(t). \quad (\text{B.12})$$

$$\check{y}_i(t) = \sum_{i_1=1}^{N_f} \dots \sum_{i_{N_f}=1}^{N_f} H_i(f_{i_1}, \dots, f_{i_{N_f}}) e^{j2\pi(f_{i_1} + \dots + f_{i_{N_f}})t}, \quad (\text{B.13})$$

$$\check{y}_1(t) = H_1(f_1)e^{j2\pi f_1 t} + H_1(f_2)e^{j2\pi f_2 t}, \quad (\text{B.14})$$

$$\check{y}_2(t) = H_2(f_1, f_1)e^{j2\pi(2f_1)t} + 2H_2(f_1, f_2)e^{j2\pi(f_1+f_2)t} + H_2(f_2, f_2)e^{j2\pi(2f_2)t}. \quad (\text{B.15})$$

$$\therefore \check{y}(t) = H_1(f_1)e^{j2\pi f_1 t} + H_1(f_2)e^{j2\pi f_2 t} + H_2(f_1, f_1)e^{j2\pi(2f_1)t} + 2H_2(f_1, f_2)e^{j2\pi(f_1+f_2)t} + H_2(f_2, f_2)e^{j2\pi(2f_2)t}, \quad (\text{B.16})$$

$$\dot{\check{y}}(t) = j2\pi(f_1 + f_2)2H_2(f_1, f_2)e^{j2\pi(f_1+f_2)t} \text{ only required}, \quad (\text{B.17})$$

$$\ddot{\check{y}}(t) = \left(j2\pi(f_1 + f_2)\right)^2 2H_2(f_1, f_2)e^{j2\pi(f_1+f_2)t} \text{ only required}. \quad (\text{B.18})$$

By substituting the frequency transforms into the model (eqn(B.1)) and equating to $2e^{j2\pi(f_1+f_2)t}$ provides the second-order function,

$$H_2(\omega_1, \omega_2) = \frac{0.5c_3 \left(H_1(\omega_1)H_1(\omega_2)j\omega_2 + H_1(\omega_1)H_1(\omega_2)j\omega_1 \right)}{\left(j(\omega_1 + \omega_2) \right)^2 - c_1}. \quad (\text{B.19})$$

Acronyms

E. coli Escherichia coli. 49, 98, 99, 103, 105, 107, 110

ABC approximate Bayesian computation. 20, 73–77, 79–81, 83–85, 87, 90, 91, 126, 128, 141, 143, 144

AR autoregressive. 25

ARMAX autoregressive moving average model with exogenous input. 25–27

ARX autoregressive model with exogenous input. 25–27

CT continuous-time. 4, 24, 40–43, 49, 51, 59–65, 67, 68, 70–73, 76, 77, 79, 83, 87, 91, 119, 123–125, 127, 129, 133, 141, 145, 147

dCTM derivative continuous-time method. 42, 49, 84, 85, 87, 89, 90

DT discrete-time. 24, 40–42, 71, 72

ERS enzymatic reaction scheme. 46, 47, 51–56, 58, 64, 65

FRO forward regression orthogonal. 36–38, 42, 63

GFP green fluorescence protein. 48, 49, 56, 58, 59, 65, 95, 96, 119, 121, 124, 128, 129

GFRF generalised frequency response function. 4, 28, 43–45, 133, 134, 137, 142

KF Kalman filter. 50

LS least squares. 33, 36, 38, 55, 62, 63, 72, 77, 81, 127

MA moving average. 25

MS model selection. 35, 36, 39, 62, 64, 143

-
- MSD** model structure detection. 28, 30, 33, 35–38, 41, 42, 58, 72, 73, 79–81, 83, 126, 128, 129, 143, 145, 147
- NARMAX** nonlinear autoregressive moving average model with exogenous input. 4, 26, 27, 36–38, 47, 72, 73
- NARX** nonlinear autoregressive model with exogenous input. 26, 27, 36–38, 41, 46, 47, 51, 59–65, 67, 68, 70, 72, 123–125, 127, 129, 133, 141, 145, 147
- NOE** nonlinear output error. 27, 33, 38, 41, 76
- OE** output error. 26, 33
- RFP** red fluorescence protein. 105, 119, 121, 124, 129, 131
- RSBP** registry of standard biological parts. 13, 14, 46, 48, 49, 97, 105, 107
- RTSS** Rauch-Tung-Striebel smoother. 50, 51
- SLS** separable least squares. 29, 31, 35, 54
- SMC** sequential Monte Carlo. 76, 77, 79–81, 85, 87, 90, 91, 126, 128, 141, 143, 144

Bibliography

- B. Ai, X. Wang, G. Liu, and L. Liu. Correlated noise in a logistic growth model. *arXiv preprint physics/0306179*, 2003.
- H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.
- H. Alper and G. Stephanopoulos. Engineering for biofuels: exploiting innate microbial capacity or importing biosynthetic potential? *Nature Reviews Microbiology*, 7(10):715–723, 2009.
- S. Anderson and V. Kadiramanathan. Modelling and identification of non-linear deterministic systems in the delta-domain. *Automatica*, 43(11):1859–1868, 2007.
- S. Anderson, N. Lepora, J. Porrill, and P. Dean. Nonlinear dynamic modeling of isometric force production in primate eye muscle. *Biomedical Engineering, IEEE Transactions on*, 57(7):1554–1567, 2010.
- E. Andrianantoandro, S. Basu, D. Karig, and R. Weiss. Synthetic biology: new engineering rules for an emerging discipline. *Molecular systems biology*, 2(1), 2006.
- A. Arkin. Setting the standard in synthetic biology. *Nature biotechnology*, 26(7):771–773, 2008.
- M. Atkinson, M. Savageau, J. Myers, and A. Ninfa. Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *escherichia coli*. *Cell*, 113(5):597–607, 2003.
- S. Atsumi, T. Hanai, and J. Liao. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature*, 451(7174):86–89, 2008.
- E. Bai. A blind approach to the hammerstein–wiener model identification. *Automatica*, 38(6):967–979, 2002.
- M. Baker. Direct protein control. *Nature Methods*, 9(5):443–447, 2012.

- T. Baldacchino, S. Anderson, and V. Kadiramanathan. Structure detection and parameter estimation for narx models in a unified em framework. *Automatica*, 48(5):857–865, 2012.
- T. Baldacchino, S. Anderson, and V. Kadiramanathan. Computational system identification for bayesian narmax modelling. *Automatica*, 49(9):2641–2651, 2013.
- Y. Bar-Shalom, X. Li, and T. Kirubarajan. *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.
- C. Barnes, D. Silk, and M. Stumpf. Bayesian design strategies for synthetic biology. *Interface focus*, 1(6):895–908, 2011.
- C. Bashor, A. Horwitz, S. Peisajovich, and W. Lim. Rewiring cells: synthetic biology as a tool to interrogate the organizational principles of living systems. *Annual review of biophysics*, 39:515, 2010.
- S. Basu, Y. Gerchman, C. Collins, F. Arnold, and R. Weiss. A synthetic multicellular system for programmed pattern formation. *Nature*, 434(7037):1130–1134, 2005.
- M. Beaumont. Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406, 2010.
- M. Beaumont, W. Zhang, and D. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- H. Berg. A physicist looks at bacterial chemotaxis. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 53, pages 1–9. Cold Spring Harbor Laboratory Press, 1988.
- S.A. Billings. *Nonlinear System Identification: NARMAX, Methods in the Time, Frequency, and Spatio-Temporal Domains*. Wiley, 2013.
- S. Billings. Identification of nonlinear systems—a survey. In *Control Theory and Applications, IEE Proceedings D*, volume 127, pages 272–285. IET, 1980.
- S. Billings and L. Aguirre. Effects of the sampling time on the dynamics and identification of nonlinear models. *International journal of Bifurcation and Chaos*, 5:1541–1556, 1995.
- S. Billings and K. Tsang. Spectral analysis for non-linear systems, part i: Parametric non-linear spectral analysis. *Mechanical Systems and Signal Processing*, 3(4):319–339, 1989a.

- S. Billings and K. Tsang. Spectral analysis for non-linear systems, part ii: Interpretation of non-linear frequency response functions. *Mechanical systems and signal processing*, 3(4):341–359, 1989b.
- S. Billings and W. Voon. A prediction-error and stepwise-regression estimation algorithm for non-linear systems. *International Journal of Control*, 44(3):803–822, 1986.
- S. Billings, M. Korenberg, and S. Chen. Identification of non-linear output-affine systems using an orthogonal least-squares algorithm. *International Journal of Systems Science*, 19(8):1559–1568, 1988.
- S. Billings, K. Tsang, and G. Tomlinson. Spectral analysis for non-linear systems, part iii: Case study examples. *Mechanical Systems and Signal Processing*, 4(1):3–21, 1990.
- C. Bishop and N. Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.
- L. Bleris, Z. Xie, D. Glass, A. Adadey, E. Sontag, and Y. Benenson. Synthetic incoherent feedforward circuits show adaptation to the amount of their genetic template. *Molecular systems biology*, 7(1), 2011.
- S. Boyd and L. Chua. Uniqueness of circuits and systems containing one nonlinearity. *Automatic Control, IEEE Transactions on*, 30(7):674–681, 1985.
- R. Breitling, F. Achcar, and E. Takano. Modeling challenges in the synthetic biology of secondary metabolism. *ACS synthetic biology*, 2013.
- J. Bruls, C. Chou, B. Haverkamp, and M. Verhaegen. Linear and non-linear system identification using separable least-squares. *European Journal of Control*, 5(1): 116–128, 1999.
- W. Burack and T. Sturgill. The activating dual phosphorylation of mapk by mek is nonprocessive. *Biochemistry*, 36(20):5929–5933, 1997.
- B. Canton, A. Labno, and D. Endy. Refinement and standardization of synthetic biological parts and devices. *Nature biotechnology*, 26(7):787–793, 2008.
- R. Carlson. The changing economics of dna synthesis. *Nature biotechnology*, 27: 1091–1094, 2009.
- R. Caspi, H. Foerster, C. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Rhee, A. Shearer, C. Tissier, et al. The metacyc database of metabolic

- pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 36(suppl 1):D623–D631, 2008.
- S. Cha and S. Srihari. On measuring the distance between histograms. *Pattern Recognition*, 35(6):1355–1370, 2002.
- D. Chandran, W. Copeland, S. Sleight, and H. Sauro. Mathematical modeling and synthetic biology. *Drug Discovery Today: Disease Models*, 5(4):299–309, 2009.
- S. Chen, S. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of control*, 50(5):1873–1896, 1989.
- R. Cheong, S. Paliwal, and A. Levchenko. Models at the single cell level. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(1):34–48, 2010.
- D. Coca and S. Billings. A direct approach to identification of nonlinear differential models from discrete data. *Mechanical systems and signal processing*, 13(5):739–755, 1999.
- C. Collins, F. Arnold, and J. Leadbetter. Directed evolution of vibrio fischeri luxR for increased sensitivity to a broad spectrum of acyl-homoserine lactones. *Molecular microbiology*, 55(3):712–723, 2005.
- A. Cornish-Bowden. *Fundamentals of enzyme kinetics*. Wiley, 2013.
- M. Covert, N. Xiao, T. Chen, and J. Karr. Integrating metabolic, transcriptional regulatory and signal transduction models in escherichia coli. *Bioinformatics*, 24(18):2044–2050, 2008.
- R. Daniel, J. Rubens, R. Sarpeshkar, and T. Lu. Synthetic analog computation in living cells. *Nature*, 2013.
- H. De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103, 2002.
- N. Draper, H. Smith, and E. Pownell. *Applied regression analysis*, volume 3. Wiley New York, 1966.
- J. Dueber, G. Wu, G. Malmirchegini, T. Moon, C. Petzold, A. Ullal, K. Prather, and J. Keasling. Synthetic protein scaffolds provide modular control over metabolic flux. *Nature biotechnology*, 27(8):753–759, 2009.

- T. Ellis, X. Wang, and J. Collins. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nature biotechnology*, 27(5):465–471, 2009.
- M. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.
- M. Elowitz, A. Levine, E. Siggia, and P. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.
- L. Endler, N. Rodriguez, N. Juty, V. Chelliah, C. Laibe, C. Li, and N. Le Novère. Designing and encoding models for synthetic biology. *Journal of The Royal Society Interface*, 6(Suppl 4):S405–S417, 2009.
- D. Endy. Foundations for engineering biology. *Nature*, 438(7067):449–453, 2005.
- K. Engel, G. Takeoka, and R. Teranishi. *Genetically modified foods*. ACS Publications, 1995.
- B. English, W. Min, A. van Oijen, K. Lee, G. Luo, H. Sun, B. Cherayil, S. Kou, and X. Xie. Ever-fluctuating single enzyme molecules: Michaelis-menten equation revisited. *Nature Chemical Biology*, 2(2):87–94, 2005.
- S. Fioretti and L. Jetto. Accurate derivative estimation from noisy data: a state-space approach. *International journal of systems science*, 20(1):33–53, 1989.
- H. Fujikawa, A. Kai, and S. Morozumi. A new logistic model for *escherichia coli* growth at constant and dynamic temperatures. *Food Microbiology*, 21(5):501–509, 2004.
- J. Garcia-Ojalvo, M. Elowitz, and S. Strogatz. Modeling a synthetic multicellular clock: repressilators coupled by quorum sensing. *Proceedings of the National Academy of Sciences of the United States of America*, 101(30):10955–10960, 2004.
- T. Gardner, C. Cantor, and J. Collins. Construction of a genetic toggle switch in *escherichia coli*. *Nature*, 403(6767):339–342, 2000.
- H. Garnier and L. Wang. *Identification of continuous-time models from sampled data*. Springer, 2008.
- H. Garnier, M. Mensler, and A. Richard. Continuous-time model identification from sampled data: implementation issues and performance evaluation. *International Journal of Control*, 76(13):1337–1357, 2003.

- D. George. Continuous nonlinear systems. Technical report, DTIC Document, 1959.
- J. Gertz, E. Siggia, and B. Cohen. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457(7226):215–218, 2008.
- F. Giri and E. Bai. *Block-oriented nonlinear system identification*. Springer, 2010.
- T. Gollisch and M. Meister. Modeling convergent on and off pathways in the early visual system. *Biological cybernetics*, 99(4-5):263–278, 2008.
- G. Golub and V. Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse problems*, 19(2):R1, 2003.
- G. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on numerical analysis*, 10(2):413–432, 1973.
- D. Goodman, G. Church, and S. Kosuri. Causes and effects of n-terminal codon bias in bacterial genes. *Science*, 342(6157):475–479, 2013.
- L. Guo and S. Billings. A modified orthogonal forward regression least-squares algorithm for system modelling from noisy regressors. *International Journal of Control*, 80(3):340–348, 2007.
- R. Haber and H. Unbehauen. Structure identification of nonlinear dynamic systems—A survey on input/output approaches. *Automatica*, 26(4):651–677, 1990.
- L. Hartwell, J. Hopfield, S. Leibler, and A. Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, 1999.
- E. Haseltine and F. Arnold. Synthetic gene circuits: design with directed evolution. *Annu. Rev. Biophys. Biomol. Struct.*, 36:1–19, 2007.
- M. Heinemann and S. Panke. Synthetic biology: putting engineering into biology. *Bioinformatics*, 22(22):2790–2799, 2006.
- S. Henriksen, A. Wills, T. Schön, and B. Ninness. Parallel implementation of particle mcmc methods on a gpu. In *System Identification*, volume 16, pages 1143–1148, 2012.
- G. Holmes, S. Anderson, G. Dixon, A. Robertson, C. Reyes-Aldasoro, S. Billings, S. Renshaw, and V. Kadiramanathan. Repelled from the wound, or randomly dispersed? reverse migration behaviour of neutrophils characterized by dynamic modelling. *Journal of The Royal Society Interface*, 9(77):3229–3239, 2012.

- S. Hoops, S. Sahle, R. Gauges, C. Lee, P., et al. Copasi: a complex pathway simulator. *Bioinformatics*, 22(24):3067–3074, 2006.
- S. Hooshangi, S. Thiberge, and R. Weiss. Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3581–3586, 2005.
- J. Houston and K. Kenworthy. In vitro-in vivo scaling of cyp kinetic data not consistent with the classical michaelis-menten model. *Drug Metabolism and Disposition*, 28(3):246–254, 2000.
- M. Hucka, A. Finney, H. Sauro, H. Bolouri, J. Doyle, H. Kitano, A. Arkin, B. Bornstein, D. Bray, A. Cornish-Bowden, et al. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.
- I. Hunter and M. Korenberg. The identification of nonlinear biological systems: Wiener and hammerstein cascade models. *Biological cybernetics*, 55(2-3):135–144, 1986.
- F. Isaacs, D. Dwyer, and J. Collins. Rna synthetic biology. *Nature biotechnology*, 24(5):545–554, 2006.
- M. Jewett, K. Calhoun, A. Voloshin, J. Wu, and J. Swartz. An integrated cell-free metabolic platform for protein production and synthetic biology. *Molecular systems biology*, 4(1), 2008.
- J. Jones and S. Billings. Recursive algorithm for computing the frequency response of a class of non-linear difference equation models. *International Journal of Control*, 50(5):1925–1940, 1989.
- R. Kalman and others. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.
- M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, et al. Kegg for linking genomes to life and the environment. *Nucleic acids research*, 36(suppl 1):D480–D484, 2008.
- G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 2008.
- R. Kass and A. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

- J. Keasling and H. Chou. Metabolic engineering delivers next-generation biofuels. *Nature biotechnology*, 26(3):298–299, 2008.
- A. Khalil and J. Collins. Synthetic biology: applications come of age. *Nature Reviews Genetics*, 11(5):367–379, 2010.
- J. Kim, K. White, and E. Winfree. Construction of an in vitro bistable circuit from synthetic transcriptional switches. *Molecular systems biology*, 2(1), 2006.
- P. Kirk, T. Thorne, and M. Stumpf. Model selection in systems and synthetic biology. *Current opinion in biotechnology*, 2013.
- M. Korenberg, S. Billings, Y. Liu, and P. McIlroy. Orthogonal parameter estimation algorithm for non-linear stochastic systems. *International Journal of Control*, 48(1):193–210, 1988.
- M. Korenberg and I. Hunter. The identification of nonlinear biological systems: Lnl cascade models. *Biological cybernetics*, 55(2-3):125–134, 1986.
- K. Krishnanathan, S. Anderson, S. Billings, and V. Kadiramanathan. A data-driven framework for identifying nonlinear dynamic models of genetic parts. *ACS synthetic biology*, 1(8):375–384, 2012.
- S. Kukreja, H. Galiana, and R. Kearney. Narmax representation and identification of ankle dynamics. *Biomedical Engineering, IEEE Transactions on*, 50(1):70–81, 2003.
- S. Kukreja, H. Galiana, and R. Kearney. A bootstrap method for structure detection of narmax models. *International Journal of Control*, 77(2):132–143, 2004.
- R. Kwok. Five hard truths for synthetic biology. *Nature*, 463(7279):288–290, 2010.
- A. Lee, C. Yau, M. Giles, A. Doucet, and C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced monte carlo methods. *Journal of Computational and Graphical Statistics*, 19(4):769–789, 2010.
- I. Leontaritis and S.A. Billings. Input-output parametric models for non-linear systems part ii: stochastic non-linear systems. *International Journal of Control*, 41(2):329–344, 1985a.
- I. Leontaritis and S. Billings. Input-output parametric models for non-linear systems part i: deterministic non-linear systems. *International journal of control*, 41(2):303–328, 1985b.

- I. Lestas, J. Paulsson, N. Ross, and G. Vinnicombe. Noise in gene regulatory networks. *Automatic Control, IEEE Transactions on*, 53(Special Issue):189–200, 2008.
- I. Lestas, G. Vinnicombe, and J. Paulsson. Fundamental limits on the suppression of molecular fluctuations. *Nature*, 467(7312):174–178, 2010.
- A. Levskaya, A. Chevalier, J. Tabor, Z. Simpson, L. Lavery, M. Levy, E. Davidson, A. Scouras, A. Ellington, E. Marcotte, and others. Synthetic biology: engineering escherichia coli to see light. *Nature*, 438(7067):441–442, 2005.
- K. Li, J. Peng, and E. Bai. A two-stage algorithm for identification of nonlinear dynamic systems. *Automatica*, 42(7):1189–1197, 2006.
- J. Liepe, H. Taylor, C. Barnes, M. Huvet, L. Bugeon, T. Thorne, J. Lamb, M. Dallman, and M. Stumpf. Calibrating spatio-temporal models of leukocyte dynamics against in vivo live-imaging data using approximate bayesian computation. *Integrative biology*, 4(3):335–345, 2012.
- J. Lin, S. Lee, H. Lee, and Y. Koo. Modeling of typical microbial cell growth in batch culture. *Biotechnology and Bioprocess Engineering*, 5(5):382–385, 2000.
- L. Ljung. *System identification*. Wiley Online Library, 1999.
- C. Lloyd, M. Halstead, and P. Nielsen. Cellml: its future, present and past. *Progress in biophysics and molecular biology*, 85(2):433–450, 2004.
- T. Lu and J. Collins. Engineered bacteriophage targeting gene networks as adjuvants for antibiotic therapy. *Proceedings of the National Academy of Sciences*, 106(12):4629–4634, 2009.
- J. MacDonald, C. Barnes, R. Kitney, P. Freemont, and G. Stan. Computational design approaches and tools for synthetic biology. *Integrative Biology*, 3(2):97–108, 2011.
- Y. Maeda and M. Sano. Regulatory dynamics of synthetic gene networks with positive feedback. *Journal of molecular biology*, 359(4):1107–1124, 2006.
- S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.
- K. Mao and S. Billings. Algorithms for minimal model structure detection in nonlinear dynamic system identification. *International journal of control*, 68(2):311–330, 1997.

- P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.
- R. McDaniel and R. Weiss. Advances in synthetic biology: on the path from prototypes to applications. *Current Opinion in Biotechnology*, 16(4):476–483, 2005.
- A. Miliadis-Argeitis, S. Summers, J. Stewart-Ornstein, I. Zuleta, D. Pincus, H. El-Samad, M. Khammash, and J. Lygeros. In silico feedback for in vivo regulation of a gene expression circuit. *Nature biotechnology*, 29(12):1114–1116, 2011.
- J. Monod. The growth of bacterial cultures. *Annual Reviews in Microbiology*, 3(1):371–394, 1949.
- D. Montgomery, E. Peck, and G. Vining. *Introduction to linear regression analysis*, volume 821. Wiley, 2012.
- D. Morgan. Cyclin-dependent kinases: engines, clocks, and microprocessors. *Annual review of cell and developmental biology*, 13(1):261–291, 1997.
- K. Müller and K. Arndt. Standardization in synthetic biology. In *Synthetic Gene Networks*, pages 23–43. Springer, 2012.
- O. Nelles. *Nonlinear system identification: from classical approaches to neural networks and fuzzy models*. Springer, 2001.
- B. Ninness. Some system identification challenges and approaches. *structure*, 1:2, 2009.
- B. Ninness and S. Henriksen. Bayesian system identification via markov chain monte carlo techniques. *Automatica*, 46(1):40–51, 2010.
- J. Niven, M. Vähäsöyrinki, M. Kauranen, R. Hardie, M. Juusola, and M. Weckström. The contribution of shaker channels to the information capacity of drosophila photoreceptors. *Nature*, 421(6923):630–634, 2003.
- B. Palsson. On the dynamics of the irreversible michaelis-menten reaction mechanism. *Chemical engineering science*, 42(3):447–458, 1987.
- R. Pearson. *Discrete-time dynamic models*. Oxford University Press, 1999.
- R. Pearson. Selecting nonlinear model structures for computer control. *Journal of process control*, 13(1):1–26, 2003.

- M. Peplow and others. Malaria drug made in yeast causes market ferment. *Nature*, 494(7436):160–161, 2013.
- V. Peterka. Bayesian system identification. *Automatica*, 17(1):41–53, 1981.
- L. Piroddi and W. Spinelli. An identification algorithm for polynomial narx models based on simulation error minimization. *International Journal of Control*, 76(17):1767–1781, 2003.
- P. Purnick and R. Weiss. The second wave of synthetic biology: from modules to systems. *Nature Reviews Molecular Cell Biology*, 10(6):410–422, 2009.
- Qiagen. Qiagen gel extraction instruction, 2010. URL <https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0CDIQFjAB&url=http%3A%2F%2Fwww.qiagen.com%2Fresources%2Fdownload.aspx%3Fid%3Df4ba2d24-8218-452c-ad6f-1b6f43194425%26lang%3Den%26ver%3D1&ei=5GWPUqu7Doy07QbkhIHACA&usg=AFQjCNFbKxQaaVix0xIbhkxizjQDx0nbLw>.
- Qiagen. Qiagen maxiprep instruction, 2012a. URL <https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&ved=0CDwQFjAD&url=http%3A%2F%2Fwww.qiagen.com%2Fresources%2Fdownload.aspx%3Fid%3D46205595-0440-459e-9d93-50eb02e5707e%26lang%3Den%26ver%3D1&ei=T2GPUv-L0oaihgeFyID4Cg&usg=AFQjCNEbJYBJ2011UgMcBo67ozmvWuUdeA>.
- Qiagen. Qiagen miniprep instruction, May 2012b. URL https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CDYQFjAA&url=http%3A%2F%2Fwww.qiagen.com%2Fresources%2Fdownload.aspx%3Fid%3D89bfa021-7310-4c0f-90e0-6a9c84f66cee%26lang%3Den%26ver%3D1&ei=916PUvXWEZTwhQfk8oHwDQ&usg=AFQjCNFwU336fBxKe0Lefx82nfd60_Vvew.
- G. Rao and H. Unbehauen. Identification of continuous-time systems. In *Control Theory and Applications, IEE Proceedings*, volume 153, pages 185–220. IET, 2006.
- J. Raser and E. O’Shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–2013, 2005.
- H. Rauch, C. Striebel, and F. Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.
- D. Ro, E. Paradise, M. Ouellet, K. Fisher, K. Newman, J. Ndungu, K. Ho, R. Eachus, T. Ham, J. Kirby, and others. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440(7086):940–943, 2006.

- N. Rosenfeld and U. Alon. Response delays and the structure of transcription networks. *Journal of molecular biology*, 329(4):645–654, 2003.
- D. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172, 1984.
- H. Salis, E. Mirsky, and C. Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature biotechnology*, 27(10):946–950, 2009.
- G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.
- N. Shaner, R. Campbell, P. Steinbach, B. Giepmans, A. Palmer, and R. Tsien. Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nature biotechnology*, 22(12):1567–1572, 2004.
- N. Shaner, P. Steinbach, and R. Tsien. A guide to choosing fluorescent proteins. *Nature methods*, 2(12):905–909, 2005.
- R. Shetty, M. Lizarazo, R. Rettberg, and T. Knight. Assembly of biobrick standard biological parts using three antibiotic assembly. *Methods Enzymol*, 498:311–326, 2011.
- Y. Shiba, E. Paradise, J. Kirby, D. Ro, and J. Keasling. Engineering of the pyruvate dehydrogenase bypass in *Saccharomyces cerevisiae* for high-level production of isoprenoids. *Metabolic Engineering*, 9(2):160–168, 2007.
- S. Shimizu-Sato, E. Huq, J. Tepperman, and P. Quail. A light-switchable gene promoter system. *Nature biotechnology*, 20(10):1041–1044, 2002.
- J. Shin. *Molecular Programming with a Transcription and Translation Cell-Free Toolbox: From Elementary Gene Circuits to Phage Synthesis*. PhD thesis, UNIVERSITY OF MINNESOTA, 2012.
- D. Siegal-Gaskins, V. Noireaux, and R. Murray. Biomolecular resource utilization in elementary cell-free gene circuits. In *The Proceedings of the IEEE American Control Conference*, to appear, 2013.
- S. Sisson, Y. Fan, and M. Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.
- S. Sisson, Y. Fan, and M. Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 106(39):16889–16889, 2009.

- T. Söderström and P. Stoica. *System identification*. Prentice-Hall, Inc., 1988.
- J. Stocker, D. Balluch, M. Gsell, H. Harms, J. Feliciano, S. Daunert, K. Malik, and J. van der Meer. Development of a set of simple bacterial biosensors for quantitative and rapid measurements of arsenite and arsenate in potable water. *Environmental science & technology*, 37(20):4743–4750, 2003.
- J. Stricker, S. Cookson, M. Bennett, W. Mather, L. Tsimring, and J. Hasty. A fast, robust and tunable synthetic gene oscillator. *Nature*, 456(7221):516–519, 2008.
- L. Sun and A. Becskei. Systems biology: The cost of feedback control. *Nature*, 467(7312):163–164, 2010.
- P. Swain, M. Elowitz, and E. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, 2002.
- W. Szybalski. In vivo and in vitro initiation of transcription. *Advances in experimental medicine and biology*, 44(1):23, 1974.
- W. Szybalski and A. Skalka. Nobel prizes and restriction enzymes. *Gene*, 4(3):181, 1978.
- A. Tamsir, J. Tabor, and C. Voigt. Robust multicellular computing using genetically encoded nor gates and chemical/wires/'. *Nature*, 469(7329):212–215, 2010.
- S. Tavaré, D. Balding, R. Griffiths, and P. Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.
- T. Tian and K. Burrage. Stochastic models for regulatory networks of the genetic toggle switch. *Proceedings of the National Academy of Sciences*, 103(22):8372–8377, 2006.
- J. Toettcher, D. Gong, W. Lim, and O. Weiner. Light-based feedback for controlling intracellular signaling dynamics. *Nature methods*, 8(10):837–839, 2011.
- T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.
- K. Tsang and S. Billings. Identification of continuous time nonlinear systems using delayed state variable filters. *International Journal of Control*, 60(2):159–180, 1994.
- J. Tucker and R. Zilinskas. The promise and perils of synthetic biology. *New Atlantis*, 12(1):25–45, 2006.

- Z. Tuza, V. Singhal, J. Kim, and R. Murray. An in silico modeling toolbox for rapid prototyping of circuits in a biomolecular 'breadboard' system.
- H. Unbehauen and G. Rao. Continuous-time approaches to system identification—A survey. *Automatica*, 26(1):23–35, 1990.
- d. Van, P. Paul, P. Bosch, V. d.K., and C. Alexander. *Modeling: identification and simulation of dynamical systems*. crc Press, 1994.
- J. van der Meer and S. Belkin. Where microbiology meets microengineering: design and applications of reporter bacteria. *Nature Reviews Microbiology*, 8(7): 511–522, 2010.
- I. Vastrik, P. D'Eustachio, E. Schmidt, G. Joshi-Tope, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, et al. Reactome: a knowledge base of biologic pathways and processes. *Genome biology*, 8(3):R39, 2007.
- V. Volterra. *Theory of functionals and of integral and integro-differential equations*. DoverPublications. com, 2005.
- B. Wang, R. Kitney, N. Joly, and M. Buck. Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology. *Nature communications*, 2:508, 2011.
- W. Weber, R. Schoenmakers, B. Keller, M. Gitzinger, T. Grau, M. Daoud-El Baba, P. Sander, and M. Fussenegger. A synthetic mammalian gene circuit reveals antituberculosis compounds. *Proceedings of the National Academy of Sciences*, 105(29):9994–9998, 2008.
- H. Wei, S. Billings, and J. Liu. Term and variable selection for non-linear system identification. *International Journal of Control*, 77(1):86–110, 2004.
- G. Welch and G. Bishop. *An introduction to the kalman filter*, 1995.
- D. Westwick and R. Kearney. Separable least squares identification of nonlinear hammerstein models: Application to stretch reflex dynamics. *Annals of Biomedical Engineering*, 29(8):707–718, 2001.
- D. Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2):122–133, 2009.