QUANTITATIVE MODELING AND ESTIMATION IN SYSTEMS BIOLOGY USING FLUORESCENT REPORTER SYSTEMS

A Dissertation

by

LOVELEENA BANSAL

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Juergen Hahn
Committee Members,	Carl Laird
	Arul Jayaraman
	Bryan Rasmussen
Head of Department,	M. Nazmul Karim

December 2013

Major Subject: Chemical Engineering

ABSTRACT

Building quantitative models of biological systems is a challenging task as these models can consist of a very large number of components with complex interactions between them and the experimental data available for model validation is often sparse and noisy. The focus in this work is on modeling and parameter estimation of biological systems that are monitored using fluorescent reporter systems.

Fluorescent reporter systems are widely used for various applications such as monitoring gene expression, protein localization and protein-protein interactions. This dissertation presents various techniques to facilitate modeling of biological systems containing fluorescent reporters with special attention given to challenges arising due to limited experimental data, simultaneous monitoring of multiple events and variability in the observed response due to phenotypic differences.

First, an inverse problem is formulated to estimate the dynamics of transcription factors, a crucial molecule that initiates the transcription process, using data of fluorescence intensity profiles obtained from a fluorescent reporter system. The resulting inverse problem is ill-conditioned and it is solved with the aid of regularization techniques. The main contribution is that, with the presented technique, any complex dynamics of transcription factors can be estimated using limited data of fluorescence measurements. The technique has been evaluated using simulated data as well as experimental data of a GFP reporter system of STAT3.

Second, an experimental design formulation is developed to facilitate the use of multiple fluorescent reporters, with overlapping emission spectra, in the same experiment. This work develops a criterion to select the fluorescent proteins for simultaneous use such that the accuracy in the estimated contributions of individual proteins to the overall observed intensity is maximized. This technique has been validated using mixtures of different *E. coli* strains which express different fluorescent proteins.

Finally, a population balance model of a cell population containing a fluorescence reporter system is developed to describe the variability in the observed fluorescence in cells. Factors such as rate of fluorescent protein formation as well as partitioning of the fluorescent protein on cell division have been taken into account to describe the dynamics of fluorescence intensity distributions in the cell populations. The model has been used to obtain preliminary hypotheses to explain the difference in response of HeLa cells containing the Tet-on expression system on stimulation by different levels doxycycline.

Thus, this work describes techniques for building robust predictive models of biological systems such as regularization for solving ill-posed estimation problems, experimental design techniques as well as using population balance modeling to model complex multi-scale dynamics. Moreover, while the examples discussed here are motivated for fluorescent reporter systems, the developed techniques can be used for different kinds of linear or non-linear dynamic biological systems.

To my parents and my brother

ACKNOWLEDGEMENTS

I would have not been able to finish my dissertation without the support of my committee members, fellow group members, my family and my friends.

My deepest gratitude is towards my adviser, Dr. Juergen Hahn who have guided me persistently not only in matters of research and technical writing but he has also been a great mentor for my professional and career development. His knowledge and technical expertise has always motivated me in my research and I had a wonderful time learning in his research group.

I am also extremely grateful to the guidance and support that I received from my committee members. I would especially like to thank Dr. Carl Laird for his insightful discussions and technical advice on all aspects of my research. I am also thankful to Dr. Arul Jayaraman for his help in improving my understanding of biological systems. He has been instrumental in my research for exploring various areas in Systems Biology through collaborative research. I am thankful to Dr. Bryan Rasmussen for being on my committee and evaluating my research all these years.

In addition, I am extremely grateful to the members in my research group for their stimulating discussions and help. I would especially like to thank Yunfei Chu and Jacky Huang for being the most motivating mentors and helpful seniors one can ever have. I also appreciate all the support from my other group members and fellow graduate students Shreya Maiti, Wei Dai, Vishal Mahindrakar, Arjun Bhadouria, Mitch Serpas, Colby Moya, Jia Kang, Yankai Cao, Gabe Hackebeil and Angelica Wong.

Most importantly, I am grateful to my parents and my brother for their constant love and support. They have always supported me in all my endeavors and instilled the personal values which have helped me in every sphere of life. Finally, I thank all my friends in College Station and in Troy for making all these years enjoyable and fun.

NOMENCLATURE

Dox	Doxycycline
EE	Elementary Effects
FI	Fluorescence Intensity
FIM	Fischer Information Matrix
FP	Fluorescent Protein
GFP	Green Fluorescent Protein
GLS	Generalized Least Squares
IBVP	Intital Boundary Value Problem
IL-6	Interleukin-6
IPDE	Integro Partial Differential Equation
LHS	Latin Hypercube Sampling
MSE	Mean Squared Error
nM	Nano Molar
O.D.	Optical Density
ODE	Ordinary Differential Equation
PBE	Population Balance Equation
PBM	Population Balance Model
RE	Relative Error
S.D.	Standard Deviation
STAT3	Signal transducer and activator of transcription 3
SVD	Singular Value Decomposition

TF	Transcription Factor
TSVD	Truncated Singular Value Decomposition

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Systems Biology aims at building quantitative models of biological processes to obtain a systems level understanding of their dynamics and the interactions between the various components that leads to the observed response (Kitano, 2002). This enables computationally simulating the changes in a biological system when it is perturbed and also testing the modifications to obtain the desired outcome. Modeling in the field of Systems Biology has ranged across various applications such as using Boolean networks and Bayesian networks to study gene regulatory systems (De Jong, 2002), ordinary differential equations to model signaling pathways (Heath and Kavraki, 2009) and stoichiometric flux analysis to understand metabolic networks (Wiechert, 2001) etc. Systems Biology further entails an integrated computational and experimental approach wherein starting from an approximate mathematical model, an iterative methodology is adopted such that the model is updated every time new experimental data is obtained (Finkelstein et al., 2004). Furthermore, hypotheses generated using the mathematical model is validated by further experimentation.

1.1 Motivation

There are numerous challenges in building reliable models of biological systems. Biological systems, such as signal transduction or metabolic pathways, are not only characterized by very large number of components but complex interactions exist between these components. These interactions can appear in the form of positive and negative feedback loops, multi-protein complexes or transcriptional controls that regulate gene expression (Aderem, 2005). Furthermore, the experimental data needed for model building or validation is often limited by sparse sampling points and mostly there are only a small number of components that can be easily measured. This situation is further complicated by a large number of unknown parameters in the models, such as rate constants in signaling pathways, which causes the parameter estimation problems in Systems Biology to be typically “underdetermined” or ill-posed.

The focus of this work is on modeling and parameter estimation of biological systems that are monitored using fluorescent reporter systems. A fluorescent reporter system involves expressing a fluorescent protein, for instance the green fluorescent protein (GFP) (Chalfie et al., 1994) , under the control of a minimal promoter such that fluorescence is observed only when the gene expression is induced by the transcriptional activator. This is done by adding the plasmid of the fluorescent reporter gene downstream of the minimal promoter so that it is also transcribed when the promoter is active.

A number of researchers in the last two decades have used fluorescence-based reporter systems for continuous and non-invasive monitoring of gene expression, protein localization and protein-protein interactions (Chalfie et al., 1994; Lippincott-Schwartz and Patterson, 2003; Roessel and Brand, 2002). The main advantage that the use of fluorescent proteins has over other approaches is that the fluorescence can be monitored in real-time, it can be easily detected using commonly-used measurement techniques, such as fluorescence microscopy, and the fluorescence can be measured without destroying the sample which allows continuous monitoring of the same sample over a period of time. This dissertation describes techniques to aid in the process of mathematical modeling and validation using fluorescent reporter systems. Special attention is given to challenges arising due to limited experimental data, large number of unknown parameters, simultaneous monitoring of multiple events, complex system dynamics and variability in the observed response due to phenotypic differences. A brief overview of the main applications described in this dissertation is given below.

1.2 Overview

Fluorescent reporter systems are widely used as indicators of transcriptional activity in cells. The first portion of the research presented in this dissertation, aims to estimate the dynamics of transcription factors (TF), i.e. the molecules that initiate the transcription process, using fluorescence intensity profiles obtained from GFP reporter systems. Determining the dynamics of transcription factor from the observed fluorescence is not straightforward as after transcription of the GFP mRNA, various other steps take place, such as its translation to form the GFP protein and fluorophore formation of GFP so that it becomes fluorescent (Roessel and Brand, 2002). Furthermore, fluorescence intensity measurements are sampled at only a few time points and are affected by a lot of variability and measurement noise. Thus, in this dissertation a regularized inverse problem has been formulated to estimate the transcription factor dynamics using limited measurements of the output fluorescence intensity.

It is also of considerable interest to monitor multiple transcription events at the same time to simultaneously investigate the behavior of several different transcription factors.

This is also crucial if interactions between different signaling pathways are to be investigated. Thus, by using two or more fluorescent protein markers with different spectral properties, multiple transcription events can be monitored at the same time. However, the emissions spectra of the different fluorescent proteins overlap and any measurement involves contributions from all the fluorescent proteins used (Dickinson et al., 2001; Lansford et al., 2001). Thus, the second part discussed in this dissertation involves investigating techniques for efficient use of multiple fluorescent markers in the same experiment.

Furthermore, there are phenotypic variations in cell populations which cause the fluorescence observed in fluorescent reporter systems to vary from cell to cell. This can be caused due to noise in gene expression, variability in signal transduction pathways (Raser and O'Shea, 2005) or physiological factors like unequal partitioning of the fluorescent protein during cell division (Hjortsø, 2006). Since most of the data from fluorescent reporter systems is obtained from cell populations, using single cell models to estimate cell physiological parameters or transcriptional dynamics with average fluorescence intensities of the sample may lead to erroneous conclusions (Hasenauer et al., 2011a). Thus, it is important to understand and separate the effect of the various factors that cause the variation in the fluorescence intensities and have an impact on the resulting fluorescence intensity distributions. In this regard, we have developed a mathematical model that describes the dynamics of the distributions of fluorescence intensity in cells containing a fluorescent reporter system.

This dissertation is organized as follows. Section 2 consists of background information and review of the techniques relevant to this work. Next, section 3 presents the formulation and solution of an inverse problem for estimating transcription factor dynamics using fluorescence profiles. Section 4 presents a novel experimental design criterion to facilitate the use of multiple fluorescent reporters in experiments. Then, Section 5 describes the technique for modeling cell populations containing a fluorescent reporter system. Finally, chapter 6 concludes with a summary of the main contributions of this work and discussion for future work. It should be pointed out that it is one of the contributions of this dissertation that real experimental data are used, wherever possible, for validating the developed techniques.

2. BACKGROUND INFORMATION

This section contains the background information and a review of the previously developed techniques used in the presented work.

2.1 Model Describing GFP Transcription and Translation

An ODE model describing transcription, translation and activation of GFP (Subramanian and Srienc, 1996) is presented below. This model consists of three ODEs which result from the component balances of the amounts of m-RNA, the non-fluorescent form of GFP, and the fluorescent form of GFP. The model had been modified (Huang et al., 2008) to take into account the constant reporter DNA levels due to stable integration of the reporter plasmid into the genomic DNA and the effect of transcription factor concentrations on the transcription rate. The resulting model is given by

$$\begin{aligned}
\frac{dm}{dt} &= S_m p \frac{C_{TF}}{c + C_{TF}} - D_m m \\
\frac{dn}{dt} &= S_n m - D_n n - S_f n \\
\frac{df}{dt} &= S_f n - D_n f
\end{aligned} \tag{2.1}$$

where C_{TF} is the concentration of the transcription factor in the nucleus, m is the mRNA concentration, n is the concentration of non-fluorescent GFP and f is the concentration of fluorescent GFP. The parameters and their constant values are: S_m is the transcription rate which is constant for a transcription factor and has a value of 373 h^{-1} for NF-kB (Huang et al., 2008) and has been re-estimated for STAT3 and C/EBP-β to be 548 h^{-1}

and 329.35 h^{-1}, respectively (Moya et al., 2009); p is the amount of DNA with a value of 5 nM; c has a value of 108 nM; D_m is the constant mRNA degradation rate that equals 0.45 h^{-1}; S_n is the translation rate and is equal to 780 h^{-1}; D_n is the protein degradation rate which equals 0.5 h^{-1}; S_f is the fluorophore formation rate which depends on the GFP variant used and it has a value of 0.347 h^{-1} for the GFP variant used in this work.

The output of the system is the mean fluorescent intensity I of a GFP reporter system and it is directly proportional to the concentration of activated fluorescent GFP in the cells

$$I = f/\Delta \tag{2.2}$$

where Δ has a value of 2.5562×10^4 nM. The initial conditions for this system are $m(0)$ = 0 nM, $n(0)$ = 0 nM, and $f(0)$ = 0 nM. Though, the mRNA levels of the fluorescence proteins may be non-zero initially but the concentrations are very low (Wang et al., 2008) and thus they can safely be assumed to be zero.

2.2 Regularization Methods for Solving Linear Inverse Problems

A main challenge in solving discrete inverse problems is that the problem can be ill-conditioned such that small perturbations in the measurements can produce large variations in the solution (Hansen, 2010; Tarantola, 2005). Regularization procedures need to be included in regression formulations to ensure stable parameter estimates are obtained. In this regard, two commonly used regularization methods, i.e., truncated singular value decomposition and Tikhonov regularization, are reviewed in this subsection.

Assuming that a linear regression model is given by

$$\tilde{\mathbf{y}} = \mathbf{H}\,\mathbf{u} + \boldsymbol{\varepsilon} \tag{2.3}$$

where $\tilde{\mathbf{y}} \in \mathcal{R}^p$ is the measurement vector, $\mathbf{u} \in \mathcal{R}^q$ is the input vector, $\mathbf{H} \in \mathcal{R}^{p\times q}$ is the transfer matrix, and $\boldsymbol{\varepsilon} \in \mathcal{R}^p$ is the measurement noise. The measurement noise is assumed to be Gaussian with zero mean and $\mathrm{rank}(\mathbf{H}) = r \leq \min\{p, q\}$.

The solution for the unknown $\mathbf{u}$ in equation (2.3) can be computed by

$$\hat{\mathbf{u}} = \mathbf{H}^{+}\tilde{\mathbf{y}} \tag{2.4}$$

where $\mathbf{H}^{+}$ is the pseudo inverse of $\mathbf{H}$. It can be evaluated as

$$\begin{aligned} \mathbf{H}^{+} &= \left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}, \qquad p \geq q \\ \mathbf{H}^{+} &= \mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{H}^{\mathrm{T}}\right)^{-1}, \qquad p < q \end{aligned} \tag{2.5}$$

where $p \geq q$ indicates an over-determined system of linear algebraic equations and the solution is obtained by using ordinary least squares. For $p < q$, the system is under-determined and the minimum-norm solution for $\mathbf{u}$ has to be calculated.

2.2.1 Truncated Singular Value Decomposition (TSVD)

The solution shown in equation (2.4) and (2.5) can be represented in an alternate form by calculating the singular value decomposition (SVD) of the transfer matrix, $\mathbf{H} = \mathbf{UWV}^{\mathrm{T}}$ where $\mathbf{U} \in \mathcal{R}^{p\times p}$, $\mathbf{V} \in \mathcal{R}^{q\times q}$ and $\mathbf{W} \in \mathcal{R}^{p\times q}$. These matrices satisfy

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_p$$

$$\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_q \tag{2.6}$$

$$\mathbf{W}_{ij} = \begin{cases} w_i & \forall\ i = j \\ 0 & \forall\ i \neq j \end{cases}$$

The diagonal entries of $\mathbf{W}$ are called the singular values of the $\mathbf{H}$ matrix and are ordered as $w_1 \geq w_2 \geq \cdots \geq w_r > 0$. Then the pseudo inverse of $\mathbf{H}$ can be calculated as

$$\mathbf{H}^{+} = \mathbf{V}\mathbf{W}^{+}\mathbf{U}^{\mathrm{T}} \tag{2.7}$$

where $\mathbf{W}^{+}$ is the pseudo inverse of $\mathbf{W}$, which can be evaluated by doing the reciprocal of all non-zero diagonal elements and transposing the matrix. Substituting equation 1(2.7) into equation (2.4), the solution can be written in the following decomposed spectral form

$$\hat{\mathbf{u}} = \sum_{i=1}^{r} \frac{\mathbf{u}_i^{T}\tilde{\mathbf{y}}}{w_i}\mathbf{v}_i \tag{2.8}$$

where $\mathbf{u}_i \in \mathcal{R}^p$ and $\mathbf{v}_i \in \mathcal{R}^q$ are columns of $\mathbf{U}$ and $\mathbf{V}$**,** respectively. The components corresponding to small singular values in equation (2.8) are responsible for large errors in the solution of discrete linear inverse problems (Hansen, 2010). Using TSVD regularization, the solution is obtained by considering only the first k components of the singular value decomposition corresponding to large singular values

$$\hat{\mathbf{u}}_w = \sum_{i=1}^{k} \frac{\mathbf{u}_i^{\mathrm{T}}\tilde{\mathbf{y}}}{w_i}\mathbf{v}_i \tag{2.9}$$

The choice of the regularization parameter k can be made on the basis of the discrete Picard condition (Hansen, 1990). According to this condition, the numerator $\mathbf{u}_i^{\mathrm{T}}\tilde{\mathbf{y}}$ should decay faster than the singular values w_i such that the overall norm of the SVD components $|\ \mathbf{u}_i^{\mathrm{T}}\tilde{\mathbf{y}}/w_i\ |$ is small. For practical applications, $|\mathbf{u}_i^{\mathrm{T}}\tilde{\mathbf{y}}|$ and the singular values are plotted for all the SVD components of the sum given in equation (2.8) and truncation parameter is chosen until the Picard condition is satisfied. Also, to obtain non-negative solutions for quantities that cannot be negative, additional non-negativity constraints need to be applied in the regularization formulation. There are few examples in literature where Truncated SVD has been implemented with non-negative constraints. These formulations range from simply setting the negative values in the estimated solutions to zero (Verkruysse et al., 2005) to mathematically more rigorous formulations involving quadratic programming problem with bounds (Villiers et al., 1999; Zhu et al., 2010).

2.2.2 Tikhonov Regularization

A least squares formulation seeks to minimize the norm of the residual between the estimated and the measured values given by $||\tilde{\mathbf{y}} - \mathbf{Hu}||_2^2$. Tikhonov regularization, also known as Ridge regression in statistics (Hastie et al., 2009), adds a regularization term to this residual, which results in the following formulation,

$$\min_{\mathbf{u}} ||\tilde{\mathbf{y}} - \mathbf{Hu}||_2^2 + \lambda\, ||\mathbf{Lu}||_2^2 \qquad (2.10)$$

Here, λ is a regularization parameter which denotes the weight of the regularization term and $\mathbf{Lu}$ is a finite difference approximation that is proportional to the first derivative of u

(Aster et al., 2005). The term $||\mathbf{Lu}||_2^2$ tries to minimize the effect of noise components by minimizing the norm of the solution. This term also aids in the solution of under-determined system of algebraic equations by decreasing the degrees of freedom. The explicit solution for this minimization problem is given by

$$\hat{\mathbf{u}}_\lambda = (\mathbf{H}^T\mathbf{H} + \lambda\, \mathbf{L}^T\mathbf{L})^{-1}(\mathbf{H}^T\tilde{\mathbf{y}}) \tag{2.11}$$

The regularization parameter can be chosen with the help of the L-Curve (Hansen, 1992) which is a plot of the norm of the residual versus the regularization term for various values of the parameter. The two norms vary monotonically with the regularization parameter with opposite trends and result in an L-shaped curve. The parameter is chosen around the corner of this L-curve to maintain a balance between the residual and the norm of the solution. This rule of thumb results from the fact that little is gained in terms of minimizing the norm of the solution by increasing the parameter λ from the one at the corner value, or with respect to minimizing the residual by decreasing λ significantly below the corner value due to the characteristic L – shape of the curve.

2.3 Optimal Experimental Design

Optimal designs are a type of experimental designs such that they allow for extraction of maximum amount of information or precision of the parameter estimates from an optimal number of experimental runs. The aim of experimental design is to reduce the variability in the parameter estimates and thus optimality of designs is evaluated based on the Fischer Information Matrix (FIM) which is proportional to the inverse of the covariance matrix of the estimates (Melas, 2005; Walter and Pronzato, 1990). Some of the statistical criterions used to evaluate the optimality of designs are the A-optimality, the D-optimality and the E-optimality criterions,

$$\begin{aligned} \Phi_A &= \min\ [tr(\mathrm{FIM}^{-1})] \\ \Phi_D &= \min\ [\det(\mathrm{FIM}^{-1})] \\ \Phi_E &= \max\ [\lambda_{\min}(\mathrm{FIM})] \end{aligned} \tag{2.12}$$

which seek to optimize the trace of the inverse of the FIM, the determinant of the inverse of the FIM and the minimum eigenvalue of the FIM respectively.

2.4 Cell Population Modeling

There are two general approaches that are widely used for modeling cell populations (see (Henson, 2003) for detailed review) - cell population balance equation (PBE) models, which are based on rigorous dynamic balances on the number of cells in a population and cell ensemble models in which a large number of individual cells are simultaneously simulated with some external interaction parameters. Each of these modeling techniques is discussed below.

2.4.1 Cell Population Balance Equation Modeling

The cell population balance equation (PBE) models are obtained by doing a dynamic balance on the number of cells using variables that characterize the intracellular state of the cells (Henson, 2003). They aim at describing the function representing the ‘distribution of cell states’. Let such a function describing the normalized distribution of states be $f(\cdot)$ and $W(\cdot)$ be the cell number concentration distribution of states. Then if the state of a cell is given by z, then $f(z,t)dz$ represents the fraction of cells with state $z \in [z+dz]$ at time t and $W(z,t)dz$ represents the number concentration of cells with state $z \in [z+dz]$. Also, these two distributions are scaled such that

$$\begin{aligned} &\int_z f(z,t)dz = 1 \\ &\int_z W(z,t)dz = N(t) \end{aligned} \tag{2.13}$$

where $N(t)$ is the total cell number concentration of cells in the population at any time t. Then, the various factors that may affect these distributions are written in terms of dynamic balances to calculate the rate of change of these distributions w.r.t. the independent variables to form the population balance model (Hjortsø, 2006).

The cell PBE models can be characterized as structured or unstructured. Unstructured PBE models are based on just one intracellular state z, whereas structured PB models have more than one intracellular state as the independent variable. However, at most, 2 or 3 states are used for formulation because PBEs involving a large number of states with multi-dimensional distribution functions become computationally intractable. The cell PBE models have been most commonly structured according to cell mass, cell age or protein or DNA content (Hjortsø, 2006; Nikos V, 2006) .

An example of the cell PBE model which is structured according to a state parameter z, which is conserved during cell division like cell mass or protein content, is given below.

$$\frac{\partial W(z,t)}{\partial t}+\frac{\partial (r(z)W(z,t))}{\partial z}=-(\Theta(z)+\Gamma(z))W(z,t)+2\int_0^\infty \Gamma(z_1)W(z_1,t)p(z,z_1)dz_1 \quad (2.14)$$

here $r(z)$ is the growth rate of the cell state z, $\Theta(t)$ and $\Gamma(t)$ are the death rate and division rate of cells respectively and $p(z,z_1)$ is the partitioning parameter such that $p(z,z_1)dz$ is the fraction of newborn cells formed with cell state between z and $z+dz$ when a cell with intensity z_1 undergoes cell division. Thus, a partial integro-differential PBE model is obtained for a conserved cell state.

2.4.2 Cell Ensemble Modeling

The cell population balance models do not incorporate the knowledge about intracellular signaling pathways in the model and the number of dependent variables that can be modeled is limited to one or two. Cell ensemble modeling, on the other hand, uses single cell models which can contain a large number of variables describing the cell state (i.e. pathway components) and the model parameters are varied to represent variability among different cells. Thus, each cell is represented by an ODE with slightly different initial conditions or reaction parameters. For instance, in one application, (Henson et al., 2002), a single cell model for yeast glycolytic oscillations was used to model an ensemble of 1000 cells which were simulated by randomizing the initial conditions of each cell in the ensemble using a Gaussian distribution.

2.5 Sensitivity Analysis

Sensitivity analysis is a powerful technique to analyze the variation in the output of a mathematical model due to the variation in the inputs or the uncertain parameters. A variety of approaches for sensitivity analysis have been developed which includes both local as well as global techniques (Butcher et al., 2004; Marino et al., 2008; Patterson et al., 2001). These techniques have widely been applied to biological systems (Marino et al., 2008; Moya et al., 2009; Saltelli et al., 2008) for analyzing the effect of the input factors such as the kinetic parameters, initial concentrations of proteins and regulatory cytokines on the responses of interest such as the activation rate of transcription factors, transcription profile of proteins or other molecules.

One of the widely used sensitivity analysis technique – Morris Method (Morris, 1991) is briefly reviewed here. It is a global sensitivity analysis technique which is convenient for identifying the important factors in a model when the number of factors is large and/or computational cost of model simulation is excessive (Patterson et al., 2001).

2.5.1 Morris Method: A Global Sensitivity Analysis Technique

The Morris method is based on calculating the elementary effects (Patterson et al., 2001) due to an input $X_i,\ i \in \{1,2,...,k\}$ on the response of the model, say Y. These elementary effects are given by,

$$EE_i = \frac{[Y(X_1, X_2,\ ...\ , X_i + \Delta,\ ...\ , X_k) - Y(X_1, X_2, ..., X_k)]}{\Delta} \qquad (2.15)$$

where EE_i is the elementary effect due to the i^{th} factor and Δ is step change in the value of the factor. Thus, an elementary effect describes the change in the value of output due to a change in the value of an input parameter while all other parameters are kept constant. The distribution of the elementary effects for each factor, $EE_i \sim F_i$, is calculated by sampling the factor values from their uncertainty range and varying only one factor at a time for calculating the difference in the numerator of equation (2.15). The sensitivity measures used are the mean and standard deviation of the elementary effects distribution for each factor. The mean of this distribution for the i^{th} factor, calculated using any r elementary effects, i.e.

$$\mu_i = \frac{1}{r}\sum_{j=1}^{r} EE_i^j \qquad (2.16)$$

gives the overall influence of the factor on the output. On the other hand, the standard deviation calculated as,

$$\sigma_i = \frac{1}{r-1}\sum_{j=1}^{r}\left(EE_i^j - \mu_i\right)^2 \qquad (2.17)$$

gives a measure of the interaction of this factor with other input factors. An efficient sampling strategy for calculating r elementary effects for each input factor has also been devised previously (Morris, 1991). Briefly, the sample space of each of the factors X_i, $i \in \{1,2,...,k\}$ is divided into an equal number of values or levels. Then, r trajectories of sample points are obtained by randomly increasing or decreasing one of the factors at a time from a base vector. In each trajectory, every factor is changed only once to obtain $k + 1$ sample points such that a total of $r(k + 1)$ points are obtained from this sampling scheme.

3. REGULARIZATION OF INVERSE PROBLEMS TO DETERMINE TRANSCRIPTION FACTOR PROFILES USING FLUORESCENT REPORTERS*

3.1 Introduction

Transcription factors (TF) are key elements of signal transduction pathways as they are involved in initiation of the transcription/translation process leading to the formation of new proteins in the cell. Thus, a quantitative description of transcription factor dynamics can aid in understanding the response of cells to external stimuli (Jothi et al., 2009; Luscombe et al., 2004). The activation of transcription factors has been conventionally monitored using protein binding methods like western blot analysis or chromatin immunoprecipitation. However, these techniques provide only qualitative or semi-quantitative data and are destructive measurement techniques, i.e., the same sample cannot be monitored continuously over time.

A number of researchers have been using fluorescence based reporter systems for continuous and non-invasive monitoring of gene expression and transcriptional activity (Chalfie et al., 1994; Roessel and Brand, 2002). Using these techniques, the underlying dynamics of transcription factors cannot be directly monitored but the fluorescence of proteins, such as the green fluorescent protein (GFP), observed from fluorescence microscopy or a fluorescent plate reader, can be used as an indicator of the activation of transcription factors. However, the relationship between the concentration of the transcription factors and the observed fluorescence is not straightforward as it involves

*Reprinted with permission from Regularization of inverse problems to determine transcription factor profiles from fluorescent reporter systems by Loveleena Bansal, Yunfei Chu, Carl Laird, Juergen Hahn, 2012. AIChE Journal 58, 3751-3762, Copyright 2012 by John Wiley and Sons.

dynamic processes dealing with transcription, translation, and post-translational modification of GFP (Roessel and Brand, 2002). A number of mathematical models describing these processes have been developed (De Jong et al., 2010; Finkenstädt et al., 2008; Leveau and Lindow, 2001; Subramanian and Srienc, 1996; Wang et al., 2008) and these models have been used for estimating the mRNA or transcription dynamics from gene expression data. For instance, in one of the studies (Ronen et al., 2002) the concentrations of the SOS transcriptional repressor for the SOS DNA repair system in E. coli were estimated but the post-translation modifications of GFP were not explicitly taken into account in the dynamic model. Other works assumed a certain nature of the dynamic profile of a compound and then estimated the parameters to characterize the profile (Finkenstädt et al., 2008; Huang et al., 2008; Wang et al., 2008). One drawback of this approach is that it restricts the functional form of the estimated profiles. This type of approach was later extended to several different functional forms (Huang et al., 2010a), however a general approach for computing the transcription dynamics has not been presented so far. Thus, in this section an inverse problem formulation is developed using which the transcription factor profiles following any dynamics can be estimated from fluorescence intensity profiles obtained from fluorescent reporter systems.

An ODE model (Huang et al., 2008) which describes transcription, translation and fluorophore formation of GFP has been used in this work (reviewed in section 2.1). The input to this ODE model is the time-dependent concentration of a transcription factor and the observed fluorescence is treated as the output. The resulting discretized inverse problem for calculating the transcription factor profiles using fluorescence intensity is also ill-conditioned. Due to this, it is an important component of this work to investigate regularization schemes which can deal with this ill-conditioned inverse problem and also filter the effect of noisy measurements on the estimated input profile. Furthermore, challenges arise since experimental data might be missing or only available at large time intervals when compared with the transcription factor dynamics. Both of these issues are also addressed in this work.

Regularization has been widely used for solving ill-conditioned inverse problems in a variety of areas including, but not limited to, electrocardiography (Dössel, 2000), geophysics (Zhdanov, 2002) and electrical impedance tomography (Borcea, 2002). A number of regularization methods for solving discrete linear inverse problems such as truncated singular value decomposition (TSVD) (Hansen, 1990), Tikhonov regularization (Aster et al., 2005; Hansen, 1992), total least squares (Fierro et al., 1997) and several iterative methods (Hansen, 2010; Vogel, 2002) exist. Among these techniques, there is no method that performs best for all types of inverse problems. Due to this, the two most commonly used methods - truncated singular value decomposition (TSVD) and Tikhonov regularization - have been used in this work for the solution of the presented inverse problem. These regularization techniques have also been implemented with non-negative constraints to obtain transcription factor concentrations which are biologically possible. This inverse problem has not been previously investigated in a discrete regularized form, thus it is not known that which regularization method is more suited for its solution. Thus, a comparison of the results obtained by the two methods is made. The reason for focusing on only these two regularization techniques is that total least squares is intended for cases where the coefficient matrix contains large perturbations (Golub et al., 2000; Shou et al., 2008) while iterative methods cater to large scale problems (Vogel, 2002). However, these two situations do not arise for the considered inverse problem dealing with computation of the transcription factor profiles.

In the next subsection, the ODE model is recast as a linear regression model and the theoretical formulation for applying regularization methods for the solution of the inverse problem is discussed. The presented technique is illustrated using two case studies in subsection 3.5. In the first case study, simulated data is used to compare the results obtained for the two regularization methods. Then the technique is applied to experimental data for estimating the profiles of the transcription factor STAT3. These experimental data are available in the form of fluorescence microscopy images obtained from the continuous stimulation of hepatocytes with 100 μg/ml of IL-6.

3.2 Procedure for Solving the Inverse Problem to Obtain Transcription Factor Profiles

In this subsection, the continuous time ODE model describing transcription and translation described in section 2.1 is discretized and expressed as a linear regression model. Using this model, regularization methods are applied to estimate the transcription factor dynamics. The problem formulation also takes into account that experimental data can often only be available at a few time points and can have missing data values.

3.2.1 Inverse Problem Formulation

The aim of this work is to compute the transcription factor profiles from fluorescence intensity measurements regardless of the specific nature of the profile. This inverse problem can be formulated as a data fitting optimization problem of the following form

$$\begin{aligned} &\min_{\hat{C}_{TF}(t)} \sum_{i=0}^{m} (\tilde{y}_i - y_i)^2 \\ &s.t. \quad y_i = g\left(\hat{C}_{TF}, \mathrm{T}_i\right) \ \ \forall\ \mathrm{T}_i = \{\mathrm{T}_0, \mathrm{T}_1, \ldots, \mathrm{T}_m\} \\ &\qquad\ \ t \in [t_0, t_n] \end{aligned} \tag{3.1}$$

where $\hat{C}_{TF}(t)$ is the continuous transcription factor profile for $t \in [t_0, t_n]$, $\tilde{y}_i$ is the discrete fluorescence intensity measurement at time T_i, y_i is the estimated intensity at time T_i using $\hat{C}_{TF}(t)$ and the model describing the transcription/translation process. This dynamic model is denoted by $y_i = g\left(\hat{C}_{TF}, \mathrm{T}_i\right)$. The range of the time interval in which measurements are available is $[\mathrm{T}_0, \mathrm{T}_m]$ and consists of $m + 1$ sampling points. The sampling step size for these measurements needs not be uniform in this formulation.

The optimization problem (3.1) is non-trivial to solve as the equality constraint consists of 3 ordinary differential equations. This continuous formulation would result in an infinite dimensional inverse problem. To avoid this, if a functional form is assumed for the profile of $\hat{C}_{TF}(t)$, it would restrict the shape of the estimated profiles. Thus, a different approach is used in this work. It is assumed that the profile for $\hat{C}_{TF}(t)$ is piecewise constant over the discretization interval and only changes between the discretization points. Furthermore, the ODE model representing the first constraint is discretized resulting in algebraic equations describing the model. Thus, the transcription factor profile is discrete and its values are estimated at each discrete time point. It should be noted that the discretization of the transcription factor profile does not have to be the same as the time points at which measurements are available.

Discretizing this particular model is aided by the fact that the model is a Hammerstein model which consists of a static input nonlinearity coupled with a linear dynamic system. Using the substitution

$$u(t) = \frac{C_{TF}(\mathrm{t})}{c + C_{TF}(\mathrm{t})} \tag{3.2}$$

eliminates the nonlinearity in the model given by equations (2.1) and (2.2) and results in a linear dynamic system. This system can be represented in the state-space form as

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t) \\ y(t) &= \mathbf{C}\mathbf{x}(t) + \varepsilon(t) \end{aligned} \tag{3.3}$$

where y is the output fluorescent intensity I and the measurement noise is denoted by $\varepsilon(t)$. The state vector and system matrices are given by

$$\mathbf{x} = [m \quad n \quad f]^{\mathrm{T}} \tag{3.4}$$

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} -D_m & 0 & 0 \\ S_n & -D_n - S_f & 0 \\ 0 & S_n & -D_n \end{bmatrix} \\ \mathbf{B} &= [S_m p \quad 0 \quad 0]^{\mathrm{T}} \\ \boldsymbol{C} &= [0 \quad 0 \quad 1/\Delta] \end{aligned} \tag{3.5}$$

As the system given by equation (3.3) is a linear dynamic system, it has a closed-form solution given by

$$y(T) = \mathbf{C} \int_0^T e^{\mathbf{A}(T-\tau)} u(\tau) d\tau \ \mathbf{B} \tag{3.6}$$

The input $u(t)$ to the system has been discretized and is assumed to be constant between two consecutive time points where discretization was performed according to the scheme shown in Figure 1. If the input and the output are sampled at different times as shown in Figure 1, then the discrete output is

$$\begin{gathered} y(T_i) = \mathbf{C} \left(\sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} e^{\mathbf{A}(T_i-\tau)} u_{j-1} \, d\tau \ + \int_{t_k}^{T_i} e^{\mathbf{A}(T_i-\tau)} u_k \, d\tau \right) \mathbf{B} = \\ \sum_{j=1}^{k} \mathbf{C} \left(\int_{t_{j-1}}^{t_j} e^{\mathbf{A}(T_i-\tau)} \, d\tau \right) \mathbf{B} \, u_{j-1} + \mathbf{C} \left(\int_{t_k}^{T_i} e^{\mathbf{A}(T_i-\tau)} \, d\tau \right) \mathbf{B} \, u_k \, , \\ \text{for} \quad t_k < T_i \leq t_{k+1} \end{gathered} \tag{3.7}$$

This solution can be represented in the form of a linear regression model as described in equation (2.3), in which the transfer matrix **H**, the output vector **y**, the input vector **u,** and the noise vector **ε** are given by

$$\mathbf{H}_{ij} = \begin{cases} \mathbf{C} \left(\int_{t_{j-1}}^{t_j} e^{\mathbf{A}(T_i-\tau)} \, d\tau \right) \mathbf{B} & t_j < T_i \\ \mathbf{C} \left(\int_{t_j}^{T_i} e^{\mathbf{A}(T_i-\tau)} \, d\tau \right) \mathbf{B} & t_{j-1} < T_i \leq t_j \\ 0 & T_i \leq t_{j-1} \end{cases} \tag{3.8}$$

$$\begin{gathered} \mathbf{y} = [y_0 \quad y_1 \quad \cdots \quad y_m]^T \\ \mathbf{u} = [u_0 \quad u_1 \quad \cdots \quad u_{n-1}]^T \\ \boldsymbol{\varepsilon} = [\varepsilon_0 \quad \varepsilon_1 \quad \cdots \quad \varepsilon_m]^T \end{gathered} \tag{3.9}$$

This formulation does not require the measurements to be available in the same time interval as the input. However, if the measurements are available in the interval $[T_0, T_m]$, the input can be calculated for the interval $[t_0, t_n]$, such that $t_n \leq T_m$. In this formulation, the discrete values of the output fluorescence intensity $\mathbf{y}$ are directly related to the input $\mathbf{u}$ which is a function of the transcription factor concentration (see equation (3.2)). The $\mathbf{H} \in \mathcal{R}^{(m+1)\times n}$ matrix usually has more columns than rows, i.e., $n > m + 1$, because experimental data are available only at a few time points but the input profile needs to be estimated at several points in time. If any data values are missing, the corresponding row can be removed from the transfer matrix and the input vector remains unchanged. The method for evaluating the integrals shown in equation (3.8) is given in the Appendix A.

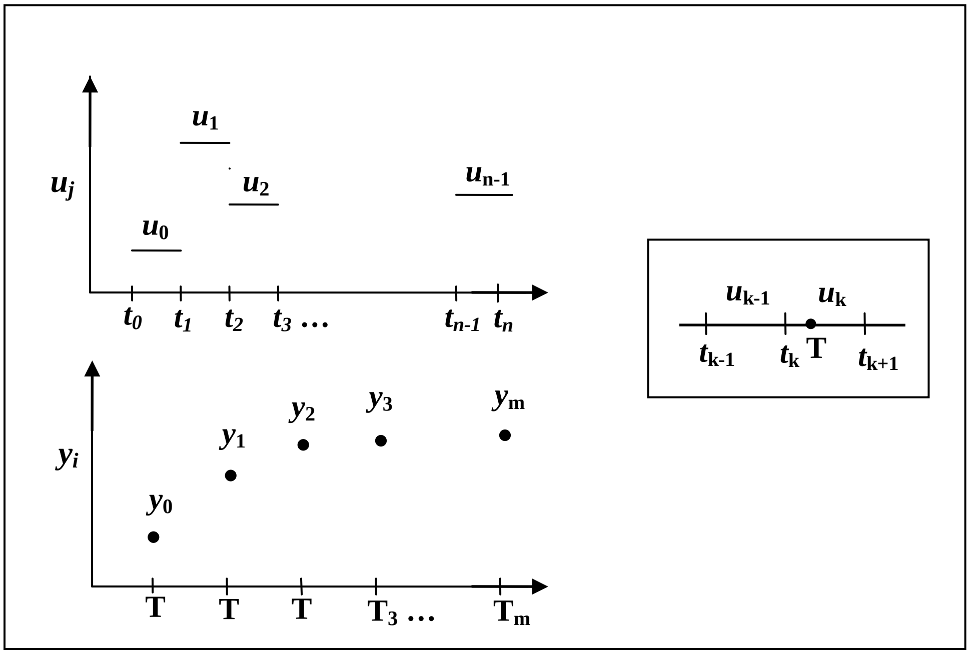


Figure 1 Discretization of the ODE model with zero-order hold for the input

The optimization problem from equation (3.1) can now be formulated as

$$\min_{\{\mathrm{C}_{TF_j}\}_{j=0}^{j=n-1}} \sum_{i=0}^{m} (\tilde{y}_i - y_i)^2$$

$$\boldsymbol{s.t.} \quad y_i = \mathbf{H}_i\mathbf{u} \;\; \forall\, i = \{0,1,\ldots,m\} \qquad (3.10)$$

$$u_j = \frac{C_{TF\,j}}{c + C_{TF\,j}} \; \forall\, j = \{0,1,\ldots,n-1\}$$

where $y_i = y(\mathrm{T}_i)$, u_j and $C_{TF\,j}$ are the constant values of the input and the transcription factor in the discrete interval $t_j \le t \le t_{j+1}$ and $\mathbf{H}_i \in \mathcal{R}^n$ is a row of the transfer matrix evaluated in equation (3.8). This formulation includes algebraic equality constraints instead of the system of ODEs which had to be solved for the original formulation in equation (3.1). This inverse problem is found to be highly ill-conditioned. Due to this, there is a need to include a regularization procedure for the solution of this inverse problem.

3.3 Application of Regularization Methods and Solution of the Inverse Problem

Discretization of the ODE model results in an under-determined linear regression model. This under-determined system forms a part of the optimization problem described by equation (3.10). This optimization problem has been transformed to solve for the unknown input $\mathbf{u}$ instead of C_{TF} such that the following formulation is obtained,

$$\min_{\mathbf{u}} \|\tilde{\mathbf{y}} - \mathbf{y}\|_2^2 \qquad (3.11)$$

$$\text{s.t.} \quad \mathbf{y} = \mathbf{H}\mathbf{u}$$

where $\tilde{\mathbf{y}} = [\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_m]^{\mathrm{T}}$, $\mathbf{y} = [y_1, y_2, \dots, y_m]^{\mathrm{T}}$ and $\mathbf{u}$ is given in equation (3.9). Transcription factor profiles are then obtained from the estimated input using the equation:

$$C_{TF}(t_j) = c\frac{u(t_j)}{1 - u(t_j)} \quad \forall\ t_0 \le t_j \le t_{n-1} \tag{3.12}$$

Regularization has been used to decrease the effect of ill-conditioning due to discretization to obtain stable solutions for the input profiles. It also decreases the effective number of parameters to be estimated and thus aids in finding the solution for this under-determined system of equations. The first derivative of the input - $\bar{\mathbf{u}}$ is regularized instead of the input $\mathbf{u}$ because it performs better for the presented inverse problem. Regularizing $\bar{\mathbf{u}}$ places a constraint on large variations in the slope of the estimated input profiles.

$$\bar{\mathbf{u}} = \mathbf{L}\,\mathbf{u} \tag{3.13}$$

where **L** is a finite element approximation matrix of the first order derivative:

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ -1 & 1 & & 0 & 0 \\ \vdots & & \ddots & \vdots & \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & & -1 & 1 \end{bmatrix}_{n\times n} \tag{3.14}$$

The transfer matrix will also need to be modified accordingly:

$$\bar{\mathbf{H}} = \mathbf{H}\,\mathbf{L}^{-1} \tag{3.15}$$

resulting in

$$\begin{aligned} &\min_{\bar{\mathbf{u}}} \|\tilde{\mathbf{y}} - \mathbf{y}\|_2^2 \\ &\text{s.t.} \quad \mathbf{y} = \bar{\mathbf{H}}\bar{\mathbf{u}} \end{aligned} \tag{3.16}$$

The regularization methods are applied to the least squares formulation given in equation above. The transcription factor concentrations are calculated from $\bar{\mathbf{u}}$ as described below.

3.3.1 Truncated SVD

The TSVD solution of the inverse problem for $\bar{\mathbf{u}}$ is evaluated by truncating the singular value decomposition of $\bar{\mathbf{H}}$ at the appropriate truncation parameter. The solution is given by

$$\bar{\mathbf{u}}_w = \sum_{i=1}^{k} \frac{\bar{\mathbf{u}}_i^{\mathrm{T}} \tilde{\mathbf{y}}}{\bar{w}_i} \bar{\mathbf{v}}_i \tag{3.17}$$

where $\bar{\mathbf{u}}_i$ and $\bar{\mathbf{v}}_i$ are columns of $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}$ and $\bar{w}_i$ are the singular values of $\bar{\mathbf{H}}$ from

$$\begin{aligned} &\bar{\mathbf{H}} = \bar{\mathbf{U}}\bar{\mathbf{W}}\bar{\mathbf{V}}^T \\ &\bar{\mathbf{W}} = \mathrm{diag}(\bar{w}_1, \cdots, \bar{w}_{\bar{r}}) \end{aligned} \tag{3.18}$$

where $\bar{r} = rank(\bar{\mathbf{H}})$. The truncation parameter is chosen from plots of $\bar{\mathbf{u}}_i^T \tilde{\mathbf{y}}$ and $\bar{w}_i$ vs i, as the maximum value of i until the Picard condition is satisfied. Then the input and the transcription factor concentrations are calculated as

$$\begin{aligned} &\mathbf{u}_w = \mathbf{L}^{-1}\bar{\mathbf{u}}_w \\ &C_{TF_w}(t_j) = c\frac{u_w(t_j)}{1 - u_w(t_j)} \quad \forall \ t_0 \le t_j \le t_{n-1} \end{aligned} \tag{3.19}$$

3.3.2 Tikhonov Regularization

The Tikhonov regularized solution for $\bar{\mathbf{u}}$ is calculated by solving

$$\min_{\bar{\mathbf{u}}} ||\tilde{\mathbf{y}} - \bar{\mathbf{H}}\bar{\mathbf{u}}||_2^2 \ + \lambda \ ||\mathbf{I}_n\bar{\mathbf{u}}||_2^2 \tag{3.20}$$

which results in

$$\bar{\mathbf{u}}_\lambda = \left(\bar{\mathbf{H}}^{\mathrm{T}}\bar{\mathbf{H}} + \lambda \, \mathbf{I}_n\right)^{-1}\left(\bar{\mathbf{H}}^{\mathrm{T}}\tilde{\mathbf{y}}\right) \tag{3.21}$$

where $\mathbf{I}_n \in \mathcal{R}^{n\times n}$ is an identity matrix. The regularization parameter λ is determined from the L-curve plotted from the values of the residual $||\tilde{\mathbf{y}} - \bar{\mathbf{H}}\bar{\mathbf{u}}||$ and the regularization term $||\mathbf{I}_n\bar{\mathbf{u}}||$ at the solution for various values of λ. The regularization parameter is then chosen close to the corner of this L-curve. The input and the transcription factor concentrations are calculated by

$$\begin{gathered} \mathbf{u}_\lambda = \ \mathbf{L}^{-1}\bar{\mathbf{u}}_\lambda \\ C_{TF\lambda}(t_j) = c\frac{u_\lambda(t_j)}{1 - \ u_\lambda(t_j)} \qquad \forall \ t_0 \le t_j \le t_{n-1} \end{gathered} \tag{3.22}$$

3.3.3 Estimation Error

Both the regularization methods, discussed in the previous sections, have been used for solving the presented inverse problem and a comparison of the results is done. The comparison is performed based upon the Relative error (RE)

$$\text{RE : } \frac{||\mathbf{\Omega}_{estimated} - \mathbf{\Omega}_{actual}||}{||\mathbf{\Omega}_{actual}||} \text{ X } 100 \tag{3.23}$$

where $\boldsymbol{\Omega}$ is the quantity to be estimated. In order to ensure a meaningful comparison, the optimal values of the regularization parameters have to be determined from the Picard plot or L-curve for each case.

3.4 Regularization with Non-Negativity Constraints

In this section, the solution of the inverse problem is evaluated by imposing additional constraints in the optimization formulation. For the estimated transcription factor profiles to be physically feasible, the concentrations at each time instant should be non-negative.

$$C_{TF}(t_j) \geq 0 \quad \forall\ t_0 \leq t_j \leq t_{n-1} \tag{3.24}$$

Since, $C_{TF}(t_j) = c\frac{u(t_j)}{1-u(t_j)}$ $\forall\ t_0 \leq t_j \leq t_{n-1}$ and $c \geq 0$ the above constraint translates to

$$\begin{aligned} u(t_j) \geq 0 \quad \forall\ t_0 \leq t_j \leq t_{n-1} \\ u(t_j) \leq 1 \quad \forall\ t_0 \leq t_j \leq t_{n-1} \end{aligned} \tag{3.25}$$

In formulation (3.16), $\bar{u}_j$ has been estimated. Given that $\bar{u}_j = u_j - u_{j-1}$, the above constraints can be written as

$$\begin{aligned} \textstyle\sum_{k=1}^{j} \bar{u}_k \geq 0 \quad \forall\, j = \{0,1,\ldots,n-1\} \\ \textstyle\sum_{k=1}^{j} \bar{u}_k \leq 1 \quad \forall\, j = \{0,1,\ldots,n-1\} \end{aligned} \tag{3.26}$$

These constraints have been included when solving the inverse problem using Truncated SVD and Tikhonov Regularization techniques. The resulting formulations are discussed below.

3.4.1 Truncated SVD with Additional Constraints

The resulting formulation for TSVD with non negativity constraints is given below.

$$\begin{aligned} &\min_{\bar{\mathbf{u}},\mathbf{y}} \|\tilde{\mathbf{y}} - \mathbf{y}\|_2^2 \\ &\text{s.t.} \quad \bar{\mathbf{u}} = \bar{\mathbf{K}}\,\mathbf{y} \\ &\qquad\;\; \mathbf{R}\,\bar{\mathbf{u}} \geq \mathbf{0} \\ &\qquad\;\; \mathbf{R}\,\bar{\mathbf{u}} \leq \mathbf{1} \end{aligned} \tag{3.27}$$

where $\mathbf{R}$ is a lower triangular matrix such that the inequality constraint incorporates the constraints in equation (3.26),

$$\mathbf{R} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{n\times n} \tag{3.28}$$

$\bar{\mathbf{K}}$ is the pseudo-inverse of the $\bar{\mathbf{H}}$ matrix which has been truncated at the appropriate truncation parameter. It is given by

$$\bar{\mathbf{K}} = \sum_{i=1}^{k} \frac{\bar{\mathbf{v}}_i \bar{\mathbf{u}}_i^T}{\bar{w}_i} \tag{3.29}$$

The truncation parameter k is chosen from the Picard condition as described before. The formulation given in (3.27) is a quadratic programming problem. It has been solved by using the solver ‘quadprog’ in MATLAB.

3.4.2 Tikhonov Regularization with Additional Constraints

$$\begin{aligned} &\min_{\bar{\mathbf{u}}} \|\tilde{\mathbf{y}} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{I}_n \bar{\mathbf{u}}\|_2^2 \\ &\text{s.t.} \quad \mathbf{y} = \bar{\mathbf{H}}\bar{\mathbf{u}} \\ &\qquad \mathbf{R}\,\bar{\mathbf{u}} \geq \mathbf{0} \\ &\qquad \mathbf{R}\,\bar{\mathbf{u}} \leq \mathbf{1} \end{aligned} \tag{3.30}$$

This above formulation for Tikhonov Regularization was also solved by using the solver ‘quadprog’ in MATLAB. The regularization parameter λ is chosen by examination of the L-curve plotted by using the solution of the above formulation.

3.5 Case Studies and Results

3.5.1 Case Study 1: Simulated Data Containing Gaussian Noise

The data used in this subsection were created by simulations, thus the real transcription factor profiles are known for this case. The next subsection describes application of the procedures to experimental data for which the transcription factor profiles are not known.

The simulated data were created by assuming a certain profile for the transcription factor concentration and then computing the fluorescence intensity profile resulting from this transcription factor profile by solving the ODE model. Gaussian noise was added to the fluorescent intensity profile to create a more realistic data set. These data are used to solve the inverse problem using both regularization methods. The estimated profiles are compared with the original assumed profiles of transcription factors to validate the accuracy of results using the relative error. As it is not possible to make comparisons for every potential input profile, the simulated data were computed for a transcription factor profile of the following form:

$$C_{TF}(t) = \mathrm{A}\,(1 - e^{-\alpha t}\, cos(t)) \tag{3.31}$$

where A and α are parameters and $t_0 \leq t \leq t_n$. This profile represents decaying oscillations which is one of the common dynamics exhibited by transcription factors. The value of A and α were assumed to be 40 and 0.3, respectively. Using this form of the transcription factor dynamics, the ODE model was simulated from t_0 = 0 to t_n = 21 hours. The sampling time for the inputs was chosen to be Δt = 0.25 hours but the fluorescence intensities data were sampled hourly. This results in a transfer matrix $\mathbf{H} \in \mathcal{R}^{21 \times 84}$.

or illustration purposes the measurements were initially simulated by adding only a small amount of random a ussian noise N(0,0.2). When no regularization has been applied for the solution of this inverse problem, the curve shown in Figure 2(a) is obtained. This solution was obtained by finding the pseudo-inverse of $\overline{\mathbf{H}}$ by using 'pinv' in MATLAB. The estimated transcription factor profile is highly erroneous with a RE of 71.5% and the estimated intensity profile, shown in Figure 2(b), completely fits the noisy data. Thus, to prevent this over-fitting and filter off the noisy components in the estimated input profile, there is a need for regularizing the solution of this inverse problem.

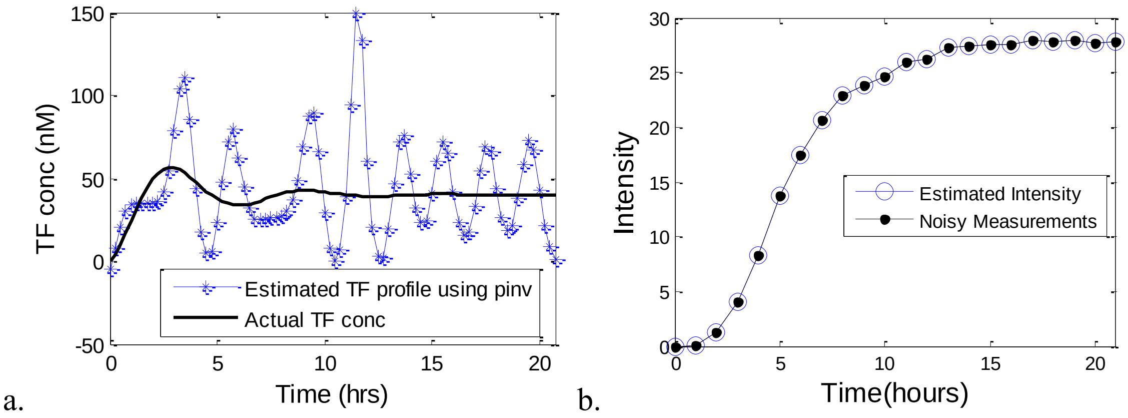


Figure 2 Solution of inverse problem using ‘pinv’ in MATLAB for measurements containing Gaussian noise N(0,0.2) a) Estimated TF profile b) Estimated intensity profile

In order to address over fitting, truncated singular value decomposition was applied to solve the inverse problem. The truncation parameter was found using the Picard condition by plotting the singular values, $\bar{\mathbf{u}}_i^{\mathrm{T}}\tilde{\mathbf{y}}$ and $|\,\bar{\mathbf{u}}_i^{\mathrm{T}}\tilde{\mathbf{y}}/\bar{w}_i\,|$ in a semi-log plot as shown in Figure 3(a). The truncation parameter was chosen to be 9 from the Picard plot as the value of the numerator was found to decay at a rate higher than the singular values until approximately the parameter 9 after which it seems to level off. The estimated profile for the optimal parameter on the basis of the Picard condition is shown in Figure 3(b) and it results in a RE of 3.68%. For comparison purposes, the estimated profiles using truncation parameters of 11 and 7, instead of the optimal 9, are shown in Figure 3(c) and Figure 3(d) respectively. It can be clearly seen that the optimal truncation parameter chosen using the Picard condition results in an estimated profile which is a better representation of the actual profile than if other values of the truncation parameter are used. If the truncation parameter is chosen larger than the optimal value, then the computed response is more oscillatory than the real response with a RE of 6.20%. Similarly, if the truncation parameter is chosen to be smaller than desired, then some of the dynamics of the input cannot be correctly reconstructed resulting in a RE of 6.87%. Also, the non-negativity constraints are redundant for this transcription factor profile as the estimated concentrations are safely above zero. So, the problem formulations without the non-negativity constraints were implemented for estimating the transcription factor profiles for the simulated data.

Tikhonov regularization is applied to the same problem for comparison purposes. The regularization parameter, i.e., the value at the corner of the L-curve was found to be approximately 150 (Figure 4a). Using this regularization parameter, the estimated profile (Figure 4b) gives a low RE of 4.22%. Thus, the results are comparable to what was found using TSVD for regularization for the case where only a small amount of noise was present in the data.

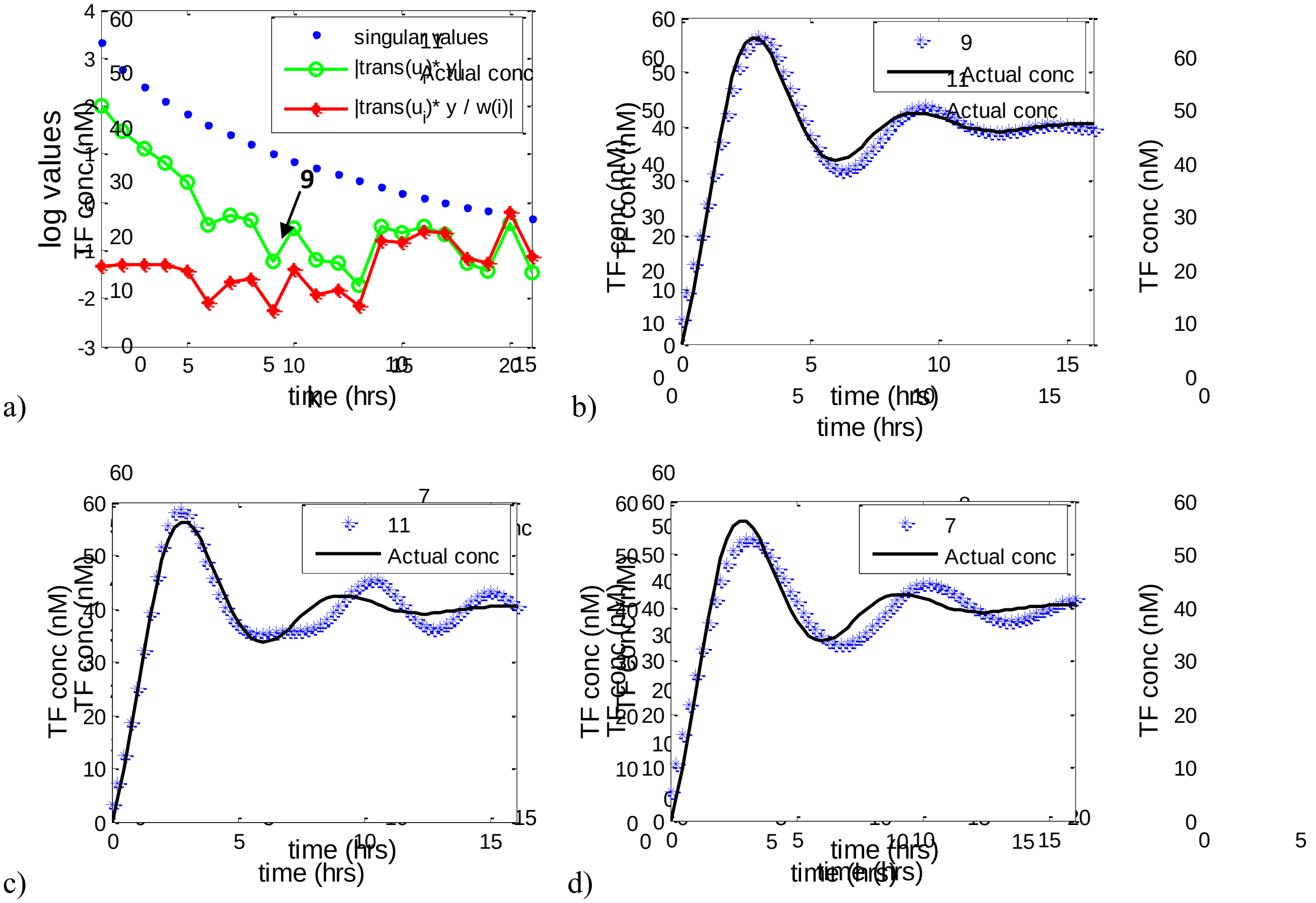


Figure 3 Solution of inverse problem by TSVD for measurements containing Gaussian noise N(0,0.2) a) Picard plot with the optimal truncation parameter of 9 b) Estimated transcription factor profile for k=9 c) Estimated profile with a truncation parameter k=11 d) Estimated profile with a truncation parameter k=7

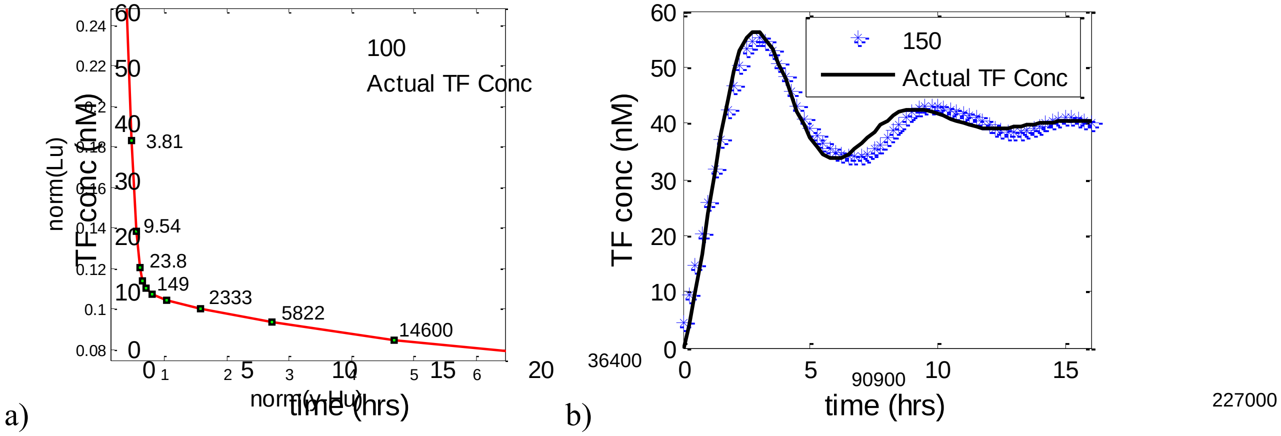


Figure 4 Solution of inverse problem by Tikhonov regularization for measurements containing Gaussian noise N(0,0.2) a) L-Curve b) Estimated TF profile using regularization parameter of 150

Solution of this inverse problem with a larger noise level has also been performed. One such simulated measurement data set containing random Gaussian noise N(0,1) is shown in Figure 5. This noise level is more realistic for biological measurements.

When applying truncated SVD to this data set, the optimal truncation parameter was observed to be 6 from the Picard plot (plot not shown). The estimated profile shown in Figure 6(a) has a relative error of 11.1%. In the estimated profile, the first peak of the profile is underestimated and the positions of the subsequent peaks are misplaced. Thus, truncated SVD does not perform as well for this example where a larger amount of noise is present in the measurements. Similar observations have been made for even larger noise levels, even though, these results are not shown here.

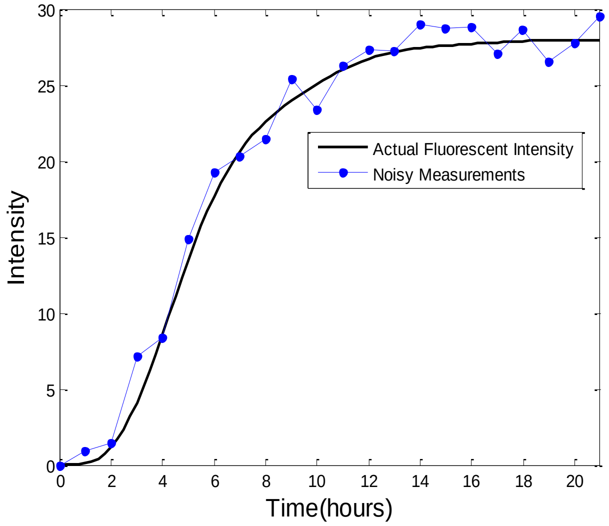


Figure 5 Simulated data containing random Gaussian noise N(0,1)

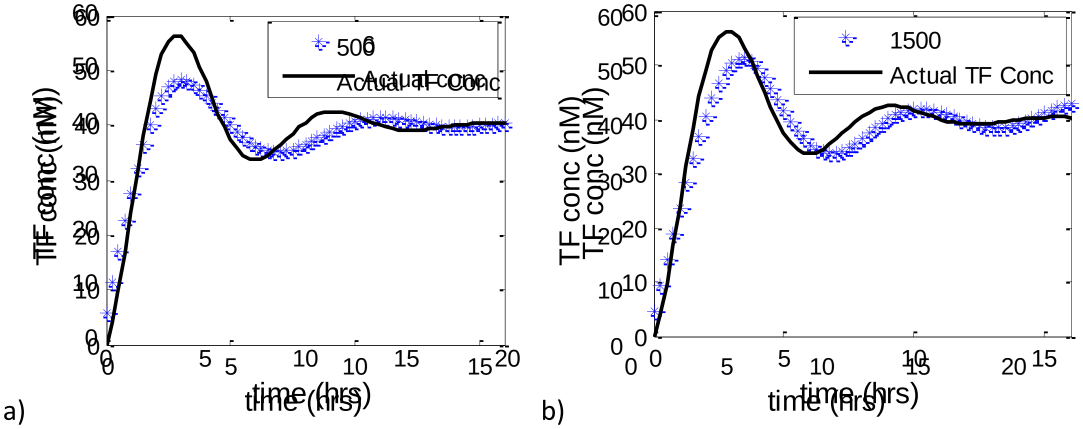


Figure 6 Estimated transcription factor profiles for simulated data containing Gaussian noise N(0,1) a) Using TSVD b) Using Tikhonov regularization

When Tikhonov regularization was used, the results are also not as good as for the low noise level case, however, the results are better than what was achieved by TSVD for this case. Using a regularization parameter of approximately 1500, obtained from the L-curve (plot not shown), the computed transcription factor profile resulted in the curve shown in Figure 6(b) which has a relative error of 8.33%.

3.5.2 Comparison between Truncated SVD and Tikhonov Regularization Using Monte Carlo Simulations

While the above shown comparisons include a significant amount of detail to explain the methods, they do not allow to draw broad conclusions as only a few specific cases were investigated. This section presents Monte Carlo simulations to provide a detailed comparison between truncated SVD and Tikhonov regularization for this inverse problem. 10,000 data sets were simulated for noise levels of N(0,0.2) and N(0,1) by the procedure described in section 3.5.1 and the inverse problem was solved for each case. The transcription factor profile was assumed to have second order dynamics. The estimated profiles do not violate the non-negativity constraints and thus these constraints were not used for solving the inverse problem for Monte Carlo simulations. The mean squared error, fitting error, squared bias and variance have been calculated from the estimated input profiles according to the following equations:

$$\begin{aligned}
&\text{Fitting Error} = tr\{E[(\mathbf{y}-\tilde{\mathbf{y}})(\mathbf{y}-\tilde{\mathbf{y}})^{\mathrm{T}}]\} \\
&\text{Mean Square Error (MSE)} = tr\left\{E\left[(\hat{\mathbf{C}}_{\mathbf{TF}}-\mathbf{C}_{\mathbf{TF}})(\hat{\mathbf{C}}_{\mathbf{TF}}-\mathbf{C}_{\mathbf{TF}})^{\mathrm{T}}\right]\right\} \\
&\text{Bias}^2 = tr\left\{(E[\hat{\mathbf{C}}_{\mathbf{TF}}]-\mathbf{C}_{\mathbf{TF}})(E[\hat{\mathbf{C}}_{\mathbf{TF}}]-\mathbf{C}_{\mathbf{TF}})^{\mathrm{T}}\right\} \\
&\text{Variance} = tr\left\{E\left[(\hat{\mathbf{C}}_{\mathbf{TF}}-E[\hat{\mathbf{C}}_{\mathbf{TF}}])(\hat{\mathbf{C}}_{\mathbf{TF}}-E[\hat{\mathbf{C}}_{\mathbf{TF}}])^{\mathrm{T}}\right]\right\}
\end{aligned} \tag{3.32}$$

where $\mathbf{C_{TF}}$ is the actual transcription factor concentration, $tr\{\cdot\}$ is the trace operator and $E[\cdot]$ is the expectation. Each data set was used to solve the inverse problem over a range of regularization parameters for each regularization method to obtain the optimal value of the parameter, i.e. the parameter that gave the lowest mean squared error. The results are shown in Table 1 for low amount of noise corresponding to N(0,0.2) and in Table 2 for higher amount of noise as represented by N(0,1). The optimal choices of regularization parameters are highlighted in bold for each regularization method. The regularization parameter for Tikhonov regularization were increased in steps of 2.5 and rounded off to the nearest integer. The inverse problem has also been solved by determining the pseudo-inverse of the transfer matrix using 'pinv' in MATLAB, i.e., without using the regularization techniques.

It can be seen from these results that Tikhonov regularization results in an approximately 26% smaller MSE than truncated SVD for data containing N(0,0.2) Gaussian noise and 10% smaller MSE for the N(0,1) noise data. Moreover, both of these methods result in significantly larger MSE when the simulated data contained a large amount of noise, as shown in Table 2. The MSE for the solution by 'pinv' is very high in both the cases and the fitting error is almost zero.

Table 1 Results from Monte Carlo simulations of 10,000 simulated data sets containing noise N(0,0.2)

	Parameter Used	**MSE**	**Fitting Error**	**Bias^2**	**Variance**
TSVD	6	699.96	1.05	639.61	61.41
	7	663.58	0.85	549.03	114.73
	8	**475.84**	**0.60**	**246.09**	**229.10**
	9	543.88	0.50	137.99	407.23
	10	846.59	0.45	119.24	727.58
Tikhonov	25	922.05	0.35	77.97	844.61
	63	499.50	0.43	87.77	411.85
	156	**350.66**	**0.54**	**147.77**	**202.48**
	391	388.49	0.74	286.64	101.53
	977	570.18	1.23	518.11	52.20
	2441	933.32	2.81	905.99	27.44
‘pinv’ solution		2.80E+05	2.26E-26	6370.488	2.74E+05

There are also some general observations that can be made about the regularization methods. For instance, the bias decreases but the variance increases when the truncation parameter is increased in truncated SVD. The reason for this is that the effective number of parameters increase and, therefore, more parameters are estimated. The same effect is caused by decreasing the regularization parameter in Tikhonov regularization. Also, the fitting errors for both the regularization methods increase with the amount of regularization as regularization tries to decrease the fit of the estimated profiles to the noisy measurements.

Table 2 Results from Monte Carlo simulations of 10,000 simulated data sets containing noise N(0,1)

	Parameter Used	**MSE**	**Fitting Error**	**Bias^2**	**Variance**
TSVD	3	8028.72	63.79	7845.07	190.78
	4	3313.80	24.60	2980.80	336.96
	5	**1754.91**	**16.73**	**965.97**	**787.18**
	6	2155.98	15.39	643.19	1515.10
	7	3431.29	14.23	524.23	2928.95
Tikhonov	375	2919.28	12.78	283.89	2641.68
	938	1839.01	14.32	504.73	1332.10
	2344	**1578.65**	**16.81**	**879.46**	**697.19**
	5859	1954.84	22.63	1576.03	374.96
	14648	3004.21	37.75	2793.19	211.24
'pinv' solution		2.81E+13	2.40E-26	4.01E+11	2.77E+13

Furthermore, the error bounds of the estimated transcription factor concentrations have been calculated to illustrate the variability in the estimated profiles. The following equation is used for calculating the standard deviation (SD) of the transcription factor concentrations at each time point,

$$e(t_i) = \left(diag\left\{E\left[\left(\hat{\mathbf{C}}_{\mathbf{TF}} - E\left[\hat{\mathbf{C}}_{\mathbf{TF}}\right]\right)\left(\hat{\mathbf{C}}_{\mathbf{TF}} - E\left[\hat{\mathbf{C}}_{\mathbf{TF}}\right]\right)^{\mathrm{T}}\right]\right\}_i\right)^{\frac{1}{2}} \quad \forall\, t_i \in [t_o, t_n] \qquad (3.33)$$

where $e(t_i)$ efers to the standard deviation for the transcription factor concentration estimated at time t_i and $diag\{\cdot\}_i$ refers to the i^{th} element of the diagonal of the matrix. The estimated profiles for the noise level – N(0,0.2) and the error bars are shown in Figure 7. The length of the error bars is $2e(t_i)$ and the expected profile and the error bars are corresponding to the regularization parameter that resulted in the lowest MSE for each regularization method. For the noise level – N(0,0.2), the standard deviations were within 4.5% and 3.7% of the expected value of the transcription factor profiles calculated using TSVD and Tikhonov regularization, respectively. This calculation excludes the errors obtained for the last few hours of the data which are significantly larger than for the rest of the profile. The reason for this there is a delay in observance of fluorescence after transcription, translation, and post-translational modification have taken place and thus the transcription factor concentrations that are computed towards the end of an experiment would largely be affected by fluorescence intensities which are occurring after an experiment has concluded. Therefore, the TF concentrations for the last few hours cannot be accurately estimated from the intensity data available for the same time range.

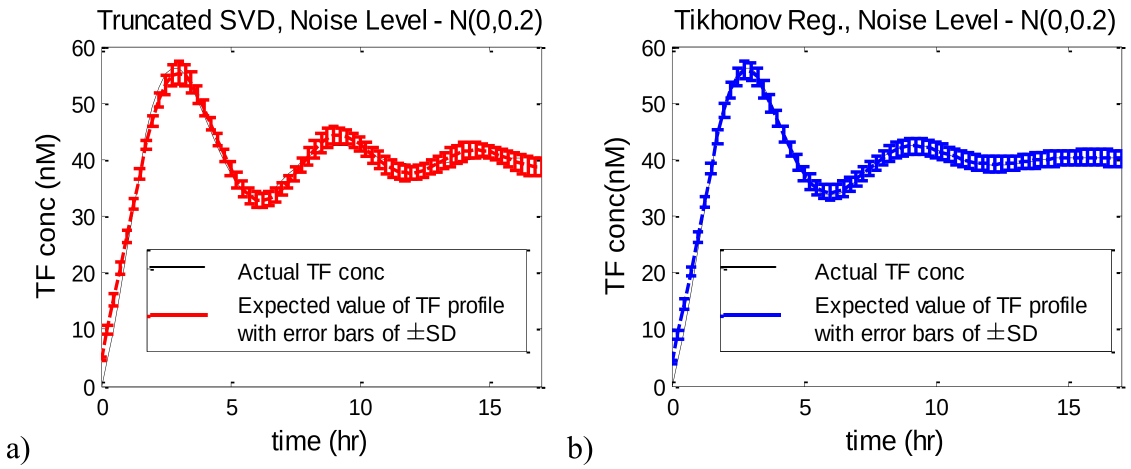


Figure 7 Expected value of TF profile for Gaussian noise level of N(0,0.2) with error bars representing ± SD using a) Truncated SVD b) Tikhonov regularization

3.5.3 Case Study 2: Application to Experimental Data Using Non-Negativity Constraints

The previous case study illustrated the effects that different choices of regularization parameters have on the solution of the inverse problem using simulated data. However, the most important test is to apply the procedure to experimental data to determine if the procedure will return satisfactory results. The experimental data are available in the form of a series of fluorescent images of a GFP reporter system (Figure 8(a)) taken during the course of the experiment. The images have been analyzed to remove noisy pixels and obtain a time-dependent mean fluorescence intensity profile (Huang et al., 2008). These data are used to estimate the dynamic profiles of transcription factor by solving the inverse problem.

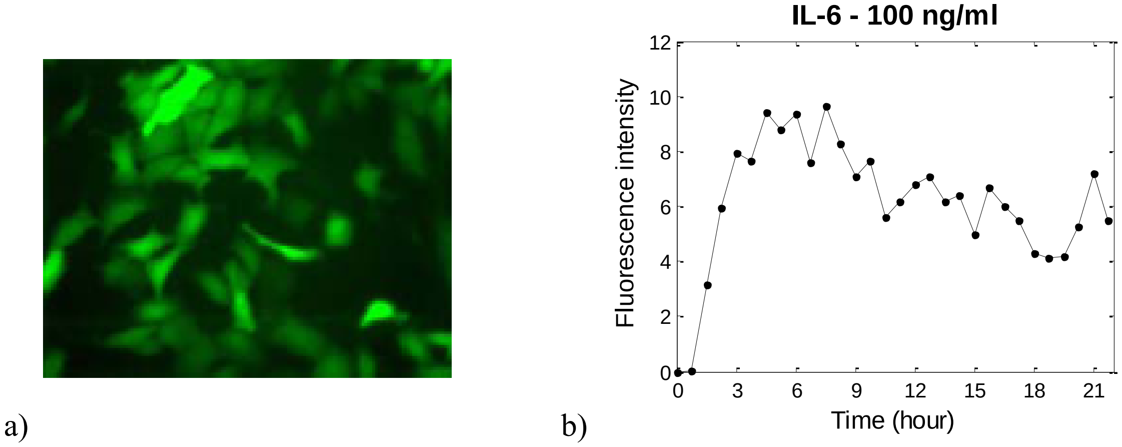


Figure 8 a) Sample image obtained from fluorescence microscopy of a GFP reporter system b) Fluorescence intensity profile obtained for IL-6-STAT3 system

For this case study, experimental data were obtained for the transcription factor STAT3 by continuously stimulating liver cells with 100ng/ml of IL-6 using a previously developed procedure (Moya et al., 2009). The fluorescent microscopy images were taken every 45min for a period of 22 hours at multiple positions in the well. The mean fluorescence intensity of the images at each time instant were calculated (Huang et al., 2008) and are shown in Figure 8(b). The shown profile was used to solve the inverse problem to obtain profiles for STAT3. It can be seen from these data that the experimental measurements contain a significant amount of noise and regularization should be applied to prevent over-fitting.

It is known that the initial dynamics of the transcription factor STAT3 shows a rapid increase followed by a steep decrease. The reason for this is that cytoplasmic STAT3 is activated and translocates to the nucleus after few minutes of stimulation with IL-6 (Kretzschmar et al., 2004; Singh et al., 2006; Watanabe et al., 2004). Thus, the input was discretized for a sampling time of 15 min to be able to infer the initial dynamics, but the sampling time was also not chosen to be too small to ensure that the inverse problem does not grow very large in size as output data were available for a time period of 22 hrs. Also, since the STAT3 concentrations cannot be negative, the two regularization techniques were applied along with non-negativity constraints and the solution was obtained by the procedure described in section 3.3. If non-negativity constraints are not used, the estimated STAT3 profiles using both TSVD and Tikhonov regularization attain negative values at a number of time points (plots not shown).

The optimal truncation parameters for this data were found to be 9 for TSVD and 350 for Tikhonov regularization from the Picard plot and L-curve, respectively. It can be seen from the estimated results shown in Figure 9 that the STAT3 profile is oscillatory in nature with a large initial peak followed by a smaller peak at around 5-6 hours and potentially another peak at around 10-11 hours. These results are consistent with Western blot data as well as simulation results of the IL-6 signal transduction pathway given in the literature (Fischer et al., 2004; Singh et al., 2006). While the solution of the transcription factor profile involving TSVD suggests similar locations of the peaks as the solution involving Tikhonov regularization, the TSVD solution seems to be more oscillatory.

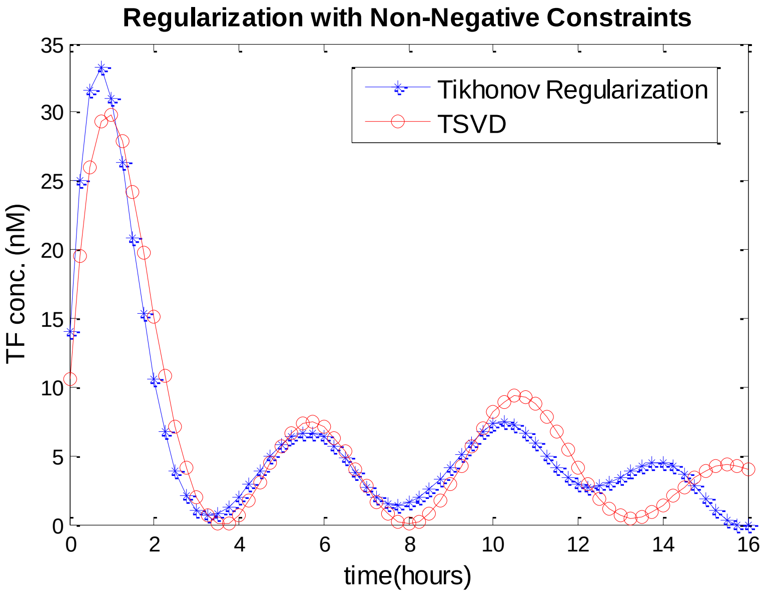


Figure 9 Estimated STAT3 profiles from Tikhonov regularization and truncated SVD

Simulation studies of the JAK-STAT pathway (Yamada et al., 2003) suggest that the ratio of the first and the second peak is approximately 5 which is more consistent with results returned by Tikhonov regularization. Moreover, the concentration profiles of the TF estimated using TSVD are almost zero at a certain time points which goes against what one would expect for this system.

3.6 Summary

This chapter presented a general method for extracting transcription factor profiles from fluorescence intensity profiles. This technique involves formulating and solving an inverse problem which directly relates the output of a GFP reporter system, i.e., the fluorescent intensity, to the input which is a function of the transcription factor concentration. The procedure used in this work places no restrictions on the shape of the input, unlike previous studies, where the transcription rates or transcription factor profiles had to be of a certain nature (Finkenstädt et al., 2008; Huang et al., 2008). This was achieved by discretizing an ODE model describing transcription, translation and fluorophore formation of GFP and then solving an inverse problem which computes the transcription factor concentration at discrete time points.

Since this inverse problem can be ill-conditioned, regularization procedures play a key role to ensure that the results are stable in the presence of measurement noise and model uncertainty. Two regularization methods, truncated SVD and Tikhonov regularization, were applied for this purpose. These regularization techniques have also been implemented along with non-negativity constraints to obtain transcription factor profiles which are physically possible. The techniques have been illustrated in two case studies where the transcription factor profiles have been computed from fluorescence intensity data using regularization. The first one considered simulated data with known inputs while the second study involved experimental data. Both methods performed satisfactorily in these case studies; however, there is an indication that Tikhonov regularization outperformed TSVD for the presented inverse problem.

4. EXPERIMENTAL DESIGN OF SYSTEMS INVOLVING MULTIPLE FLUORESCENT PROTEIN REPORTERS*

4.1 Introduction

Fluorescent proteins are now regularly used to monitor a variety of aspects of biological systems. There are also a number of fluorescent proteins that are commercially available (Nowotschin et al., 2009; Shaner et al., 2005) which can be simultaneously used for profiling of biological systems. The main advantage of concurrently using multiple fluorescent reporters, with different emissions spectra, is that it is possible to simultaneously monitor different components of a system or their interactions. For instance, a number of researchers have used multiple fluorescent reporters and multispectral imaging techniques to monitor complex protein-protein interactions (Hu and Kerppola, 2003; Waadt et al., 2008), protein and cellular movements (Hiraoka et al., 2002; Hoffman, 2005), transcriptional regulation due to multiple promoters (Cox et al., 2010) and for understanding morphological developments in in-vitro or in in-vivo imaging (Chen et al., 2012; Hoffman, 2005; Nowotschin et al., 2009). In addition, labeling with multiple fluorescent markers has also been used for determining the compositions of different bacterial species in biofilms (Cowan et al., 2000; Ma and Bryers, 2010). However, the combined use of multiple fluorescent proteins is challenging due to the fact that the emission spectra of different proteins overlap and any measurement involves contributions from all the fluorescent proteins. Thus, extracting

* Reprinted with permission from Experimental design of systems involving multiple fluorescent protein reporters by Loveleena Bansal, Randall Nelson, Eric Yang, Arul Jayaraman, Juergen Hahn, 2013. Chemical Engineering Science 101, 191-198, Copyright 2013 by Elsevier.

the contribution of individual types of fluorescent proteins from measurements becomes very difficult.

A number of "unmixing" algorithms have been discussed in the literature to resolve the overlapping emission spectra of fluorescent reporters in measurements from samples containing two or more reporters at the same time. If only two reporters are present in a sample, one can choose the reporters with minimum overlap in their emission or excitation spectra and use appropriate filter sets to separately measure the intensity of the reporters (Cowan et al., 2000; Shaner et al., 2005). In a large number of applications involving three or more fluorescent protein reporters, the spectral contributions of the individual reporters are distinguished using imaging spectroscopy aided with a mathematical linear unmixing formulation in which the fluorescence intensity of each pixel in an image is assumed to be a linear combination of the intensities due to individual reporters (Dickinson et al., 2001; Lansford et al., 2001; Zimmermann et al., 2003). This technique has been implemented to distinguish the spectra of up to even seven different fluorescent dyes (Tsurui, 2000). Additionally, flow cytometry has widely been used for high-throughput analysis at the single cell level to measure a very large number of fluorescence signals by using up to 30-40 optical filters for excitation blocking, spectral separation and transmission of narrow bands (Grégori et al., 2012; Perfetto et al., 2004). The overlapping excitation and emission spectra of fluorescent reporters has entailed the use of intricate and expensive experimental setups for their spectral resolution and remains a major challenge in multicolor applications.

In this chapter, a mathematical approach is discussed for selecting the fluorescent proteins to use together for multiple labeling applications in order to reconstruct the contribution of individual proteins to the overall measurements. Two main tasks for selection of the fluorescent reporters have been addressed: (1) solution of an unmixing problem to determine the contribution of individual proteins to the overall intensity of a sample containing two or more proteins, and (2) making use of this formulation to perform model based experimental design such that the accuracy in the estimates of the contribution of individual proteins to the overall observed fluorescence is maximized. For the former task, the overall fluorescence intensity of a sample is measured using a plate reader and it is assumed to be a linear superposition of the intensities of the individual proteins present in it. The goal in the latter is to make the decision of which different fluorescent proteins should be used in an experiment amongst the available proteins based on the D-optimality design criterion (Hinkelmann, 2012; Melas, 2005). The developed techniques have been validated using experimental data from mixtures of different *E. coli* strains where each type of *E. coli* expressed a different fluorescent protein.

4.2 Methods

For the purpose of this study, *E. coli* strains, suspended in a clear media and expressing different fluorescent proteins, were mixed together to create mixtures containing up to three different fluorescent proteins, similar to what is commonly done in the field of biofilm analysis (Ma and Bryers, 2010).

4.2.1 Bacterial Strains, Plasmids and Cell Culture

The *E. coli* strains, the fluorescent protein plasmids used and their sources are listed in Table 3. For cell culture, the *E. coli* strains containing the corresponding plasmid were taken from glycerol stocks and streaked on LB agar plates supplemented with an appropriate antibiotic to maintain the plasmid (150 μg/mL erythromycin for the GFP-expressing strain and 100 μg/mL ampicillin for RFP, CFP, and YFP-expressing strains). The plates were incubated overnight at 37°C. *E. coli* from the fresh LB agar plates were cultured overnight for 14 hours in 25 ml of tryptone broth media (TB; 10 g tryptone/L H_2O and 8 g NaCl/L H_2O) with the appropriate antibiotic at 37°C with shaking. Then, a part of the cell culture (15 mL) was centrifuged at 3000 RPM for 7 minutes. After visually confirming the presence of a cell pellet, the supernatant media was removed and the bacteria were resuspended in 10 mL of clear chemotaxis buffer (CB; 1× phosphate-buffered saline, 0.1 mM EDTA (pH- 8.0), 0.01 mM L-methionine, and 10 mM DL-lactate) and diluted to different concentrations based on OD_{600} measurements.

4.2.2 Mixtures of Different *E. coli* Strains

The *E. coli* strains expressing different fluorescent proteins were mixed in different ratios in a 96-well plate. Several mixtures were made containing up to 3 different fluorescent protein-expressing *E. coli* strains. The final volume of the mixture used in each well was 150 μl. Thus as an example for making a 1:1:1 mixture (by volume) of *E. coli* RP437(pCM18), *E. coli* TG1(pAmCyan) and *E. coli* TG1(pZsYellow) from their individual cultures, 50 μl of each of these strain cultures were mixed together. The well plate was mixed mechanically to ensure proper mixing of the strains and their emission spectra were measured. The emission spectra of individual fluorescent proteins were also recorded for 150 µl of its culture at different concentrations (or optical densities) which were used for creating the mixtures.

Table 3 *E. coli* strains and fluorescent protein plasmids

***E. coli* Strain**	**Plasmid**	**Plasmid Description**
E. coli RP437 (Mao et al., 2003)	pCM18 (Hansen et al., 2001)	Green Fluorescent Protein (GFP)-expression plasmid
E. coli TG1 (Stratagene, La Jolla, CA)	pDsRed-Express (Clontech, CA)	Red Fluorescent Protein (RFP)-expression plasmid
	pAmCyan (Clontech, CA)	Cyan Fluorescent Protein (CFP)-expression plasmid
	pZsYellow (Clontech, CA)	Yellow Fluorescent Protein (YFP)-expression plasmid

4.2.3 Fluorescence Intensity Measurements Using a Plate Reader

The fluorescence intensity measurements of individual *E. coli* strains and their mixtures were taken using a plate reader. The reason for using a plate reader is that it allows the measurement of emission intensity at various wavelengths which is not easily possible using fluorescence microscopy.

Table 4 Maximum excitation and emission wavelengths of the four fluorescent proteins

Plasmid	**Excitation maximum (nm)**	**Emission maximum (nm)**
pAmCyan (CFP)	458	489
pCM18 (GFP)	490	520
pZsYellow (YFP)	529	539
pDsRed-express (RFP)	554	586

The maximum excitation and maximum emission wavelengths for each fluorescent protein used are given in Table 4. The excitation wavelengths for the mixtures of fluorescent proteins were chosen such that all the fluorescent proteins in the mixture are excited at the same time. It was also taken into account that the excitation intensity should not overlap with the emission spectra of the proteins. For example, for a mixture containing *E. coli* RP437(pCM18) and *E. coli* TG1(pAmCyan), the excitation wavelength of 430 nm was used at which both GFP and CFP are excited (Figure 10a) and the intensity of the excitation beam at 430 nm minimally overlaps with the emission spectra of these proteins (Figure 10b). The emission spectrums of the individual GFP and CFP cultures, for this case, were also recorded at 430 nm. The Gemini^{TM} EM Fluorescence Microplate Reader by Molecular Devices was used for taking all the measurements. The measurements were taken using the 'bottom read' option in the plate reader and the emission intensities of the mixtures were measured at multiple emission wavelengths. The emission intensities of CB media were also recorded to use as control and subtracted from intensity measurements of fluorescent proteins and their mixtures.

All the measurements were obtained as triplicates and averaged to reduce the measurement noise.

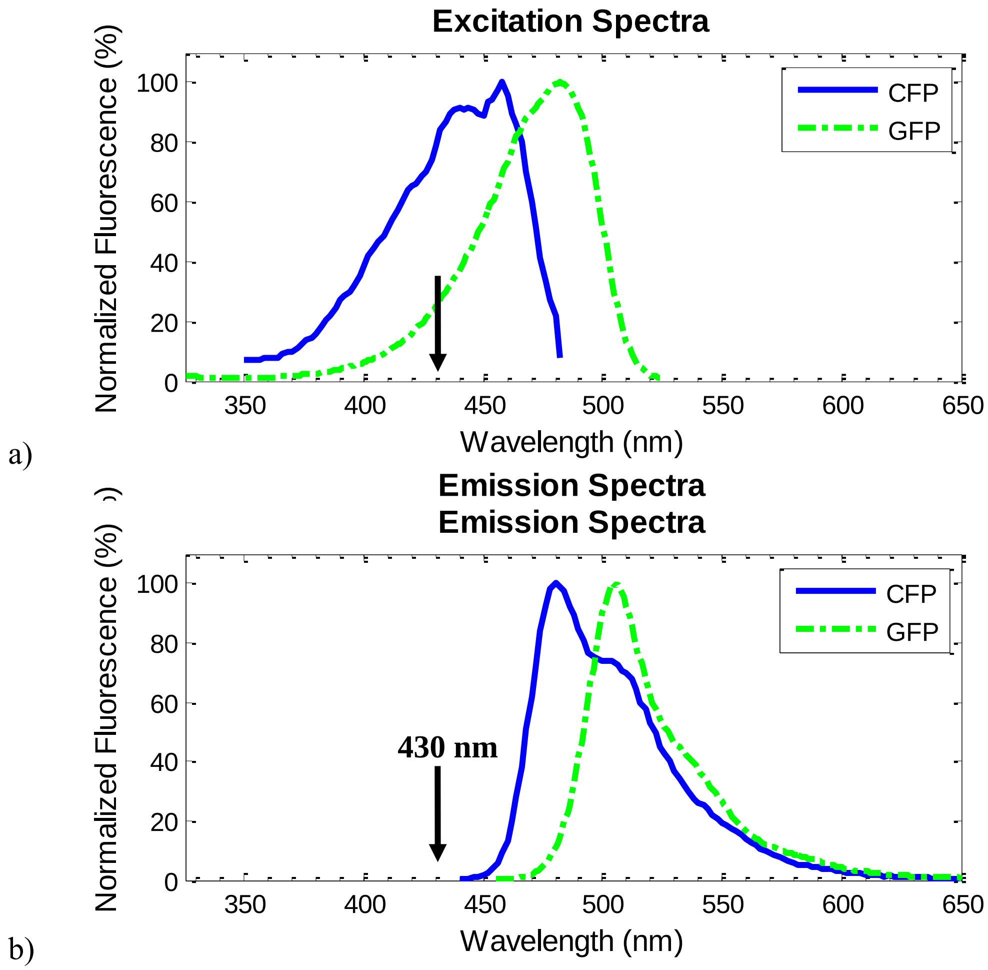


Figure 10 Illustration for choosing the excitation wavelength for a mixture containing CFP and GFP (data obtained from an online database (Biosciences, 2000) for CFP and GFP variants that have maximum emission and excitation wavelengths reasonably close to pAmCyan and pCM18 protein plasmids, respectively)

4.2.4 Linear Unmixing of Fluorescent Strains

As described in the previous section, the emission spectra of mixtures as well as of individual strains expressing a single fluorescent protein have been measured. These measurements were used to extract the intensity contribution of individual fluorescent proteins from the overall intensity measurement. The procedure for this step is described below.

Let the emissions spectra of individual proteins be described by a function $g_j(x)$, where x is the emission wavelength, $g_j(x)$ is the fluorescence of the *j*-th protein emitted at this wavelength, and $j = 1,...,k$ is the index referring to different fluorescent proteins. Then, for different emission wavelengths:

$$g_j(x) = g_j(x_i) \; for \; x_i \le x < x_{i+1}, \; 0 < i \le n \tag{4.1}$$

where n is the number of discrete wavelengths at which the emission intensity is measured. For mixtures containing multiple fluorescent proteins, the emission intensity can be represented as a linear superposition of the intensities of individual proteins, such that

$$y(x) = \theta_1 g_1(x) + \theta_2 g_2(x) + \cdots + \theta_k g_k(x) + e(x) \tag{4.2}$$

where $y(x)$ is the emission intensity of the mixture and $e(x)$ is the measurement noise and, without loss of generality, it is assumed that the noise is Gaussian distributed as $e(x) \sim N(0, \sigma^2)$. The parameters θ_j for $j = 1,...,k$ represent the contribution of the emission intensity of the j-th fluorescent protein to the mixture intensity.

For different emission wavelengths, equation (4.2) can be expanded as,

$$\begin{aligned}
&y(x_1) = \theta_1 g_1(x_1) + \theta_2 g_2(x_1) + \cdots + \theta_k g_k(x_1) + e(x_1) \\
&y(x_2) = \theta_1 g_1(x_2) + \theta_2 g_2(x_2) + \cdots + \theta_k g_k(x_2) + e(x_2) \\
&\vdots \\
&y(x_n) = \theta_1 g_1(x_n) + \theta_2 g_2(x_n) + \cdots + \theta_k g_k(x_n) + e(x_n)
\end{aligned} \tag{4.3}$$

which can be rewritten as

$$\mathrm{Y}(x) = \mathrm{G}(x)\Theta + \mathrm{E}(x) \tag{4.4}$$

using vector and matrix notations. Since the set of fluorescent proteins that are used in an experiment are always known, the quantity $\mathrm{G}(x)$, i.e. the emission spectra of the individual proteins, is also known. Similarly, the set of wavelengths $x_i,\ i = 1...n$, at which the fluorescence intensity is measured can be chosen prior to an experiment. The problem of computing the contribution of the different fluorescent proteins to the overall intensity is given by computing Θ from equation (4.4) from the set of mixture intensity measurements. The resulting optimization problem is,

$$\begin{aligned}
&\min_{\Theta} \sum_{i=1}^{n} \left(y(x_i) - \hat{y}(x_i) \right)^2 \\
&\textit{subject to} \\
&\hat{y}(x_1) = \theta_1 g_1(x_1) + \theta_2 g_2(x_1) + \cdots + \theta_k g_k(x_1) \\
&\hat{y}(x_2) = \theta_1 g_1(x_2) + \theta_2 g_2(x_2) + \cdots + \theta_k g_k(x_2) \\
&\vdots \\
&\hat{y}(x_n) = \theta_1 g_1(x_n) + \theta_2 g_2(x_n) + \cdots + \theta_k g_k(x_n)
\end{aligned} \tag{4.5}$$

where $\hat{y}(x_i)$ for $0 < i \le n$ are the measured emission intensities for a mixture of fluorescent proteins. The least squares solution for the optimization problem in equation (4.5) can be written as,

$$\Theta = \left(\mathrm{G}^T(x)\mathrm{G}(x)\right)^{-1}\mathrm{G}^T(x)\mathrm{Y}(x) \tag{4.6}$$

If the noise term is Gaussian distributed as $e(x) \sim N(0,\sigma^2)$ then Θ will follow the distribution

$$\Theta \sim N(0,\sigma^2(\mathrm{G}^T\mathrm{G})^{-1}) \tag{4.7}$$

where σ^2 is the variance of the measurement noise and G is the emission spectrum matrix $\mathrm{G}(x)$. In this formulation, it is assumed that the error in the emission spectrum matrix G is negligible. However, since the emission spectra of individual fluorescent proteins are also measured using the plate reader, there is some uncertainty in these measured values due to inevitable factors such as photon shot noise, detector noise (Brukilacchio, 2003; Neher and Neher, 2004) or background noise (Zimmermann et al., 2003) etc. However, analysis of the effects of these noise factors on the unmixing formulation is beyond the scope of this work. Furthermore, the solution of this optimization problem forms the basis for the experiment design procedure that is discussed in the next subsection.

4.2.5 Experimental Design Criterion to Select the Set of Fluorescent Proteins

The aim of the presented experimental design formulation is to determine a set of fluorescent proteins to use such that the contributions of the different types of proteins can be estimated as accurately as possible. In this work, the optimal design is obtained by minimizing the D-optimal design criterion which is based on the determinant of the Fischer information matrix (Melas, 2005; Walter and Pronzato, 1990):

$$[\det(\mathrm{G}^T\mathrm{G})]^{-1} \tag{4.8}$$

where $\det(\bullet)$ denotes the matrix determinant operator. A larger value of $\det(\mathrm{G}^T\mathrm{G})$ results in smaller volume of the confidence region for the estimated value of Θ thereby resulting in more accurate estimates (Walter and Pronzato, 1990). Based upon this, the following optimization problem is formulated,

$$\begin{aligned}
&\max_{j_1,\cdots,j_k} \det\left(\mathrm{G}^{\mathrm{T}}\mathrm{G}\right) \\
&\textit{subject to} \\
&\mathrm{G} = \left[\mathbf{g}_{j_1} \quad \mathbf{g}_{j_2} \quad \cdots \quad \mathbf{g}_{j_k}\right] \\
&j_l \in \{1,2,...,m\}, \quad \textit{for} \quad l = 1,2,..,k \\
&\mathbf{g}_j = \left[g_j(x_1) \quad g_j(x_2) \quad \cdots \quad g_j(x_n)\right]^{\mathrm{T}}
\end{aligned} \tag{4.9}$$

where the matrix G is affected by the fluorescent proteins chosen for an experiment, particularly their emission spectra $\mathbf{g}_j$ as well as the wavelengths at which measurements are taken, i.e. x_i. Since the wavelengths at which the emission spectra are measured are generally pre-determined, the number of rows of G as well as all x_i, $i = 1, 2,..., n$ are fixed. Also, the number of fluorescent proteins that need to be used is known as one generally knows beforehand the number of different events that need to be monitored. This results in a value of k that is fixed. What remains is to choose k proteins j_l, $l = 1, 2, ..., k$ from a set of m possible proteins that are available. This results in a mixed-integer programming (MIP) problem.

The number of available proteins (m) and the events that need to be monitored (k) are not large for most applications, e.g., determining the three best proteins to use out of 8-10 different proteins that are available in a lab. In that case, it is possible to solve the MIP by exhaustive search and testing all possible combinations. This involves computing the value of $\det(\mathrm{G}^{\mathrm{T}}\mathrm{G})$ for all possible combinations of k fluorescent proteins out of the available m proteins, where the number of such combinations is given by the binomial coefficient C_k^m, and choosing the combination with the largest value of $\det(\mathrm{G}^{\mathrm{T}}\mathrm{G})$. However, for problems involving larger number of fluorescent proteins, e.g., choosing the best combination of six proteins out of 30 commercially available ones, such a brute force approach may not be practical. Instead a simple procedure based upon forward selection can be used to find an approximate solution of this MIP (Chu and Hahn, 2012). Such a forward selection procedure has the following form,

Step 0 (Initiation). Set the number of proteins selected to one, i.e., $l = 1$ (4.10)

Step 1 (Selection). Select the protein indexed by j_l which is determined by

$$j_l = \arg\max_j \ \mathbf{g}_j^{\mathrm{T}} \mathbf{g}_j \tag{4.11}$$

Step 2 (Projection). Let $\mathbf{g}_j = \mathbf{g}_j - \dfrac{\mathbf{g}_j^{\mathrm{T}} \mathbf{g}_{j_l}}{\mathbf{g}_{j_l}^{\mathrm{T}} \mathbf{g}_{j_l}} \mathbf{g}_{j_l}$ (4.12)

Step 3 (Stopping test). If $l < k$, return to Step 1 with $l = l + 1$ (4.13)

In the first step, the algorithm selects the fluorescent protein for which the square of the 2-norm of its emission spectrum, i.e., $\| \mathbf{g}_j \|_2^2$ is maximized. In step 2, the vectors of the unselected proteins are projected on to the space orthogonal to that spanned by the vector of the previously selected protein in order to remove the protein's effect on the output covered by the already selected proteins. This procedure is repeated until k

proteins have been selected. This procedure will result in the selection of fluorescent proteins which have minimum overlap in their emission spectra.

4.3 Results

The results for extracting the contribution of individual proteins to overall mixture intensities are presented here for mixtures containing two or three different fluorescent protein-expressing bacterial strains.

4.3.1 Extracting the Contribution of Individual Proteins from Mixture Intensities

Bacterial strains expressing different fluorescent proteins were mixed in a 96-well plate as described in section 4.2.2. The emission spectra of the mixtures and the individual proteins were then measured using a fluorescent plate reader. Several such mixtures were investigated which contained different number and types of fluorescent proteins. Furthermore, a number of data sets of mixtures of the same proteins were obtained by using proteins at different concentrations (as measured using OD_{600} values).

The emission spectra obtained for some of the mixtures containing two fluorescent proteins are illustrated in Figure 11. In each of the sub-plots, the emission intensity of the individual proteins (at the corresponding optical densities) and of the mixture obtained by mixing those proteins in the specified ratio is plotted. The emission intensity of CB media measured at the same excitation and emission wavelengths has been subtracted from all these spectra. For these measurements, the excitation wavelength was chosen such that all the fluorescent proteins in the mixture are excited at the same time. This led to the emission intensity of red and yellow fluorescent proteins

(FPs) to be lower compared to the intensity of green or cyan FPs (Figure 11b and Figure 11d), however this difference did not cause observable inaccuracies in the results. The emission spectra of the individual proteins to be used in the G matrix were also obtained at the corresponding excitation wavelength for solving the unmixing problem.

As the ratios in which the different fluorescent protein-expressing bacterial strains are mixed are known, the contributions of individual proteins to the overall mixture spectra, i.e. the theoretical values of θ_j's, are known. Thus, as an example, for a 1:2 mixture (by volume) of red and yellow fluorescent protein cultures (Figure 11a), the theoretical values of θ_j are 0.33 and 0.67 for RFP and YFP, respectively. The linear unmixing formulation discussed in section 4.2.4 is then used to estimate the values of θ_j to validate the assumption of linear superposition of the spectra of individual proteins. The relative error between the theoretical and the estimated θ_j values is used to evaluate the accuracy of the presented formulation:

$$\text{Relative Error} = \left(\frac{\theta_{Actual} - \theta_{Estimated}}{\theta_{Actual}} \right) \times 100 \tag{4.14}$$

A summary of the overall relative errors for estimating the θ_j's for different 2-protein mixtures is given in Table 5. In these results, the largest mean relative error was approximately 14% and most of these errors are below 10%.

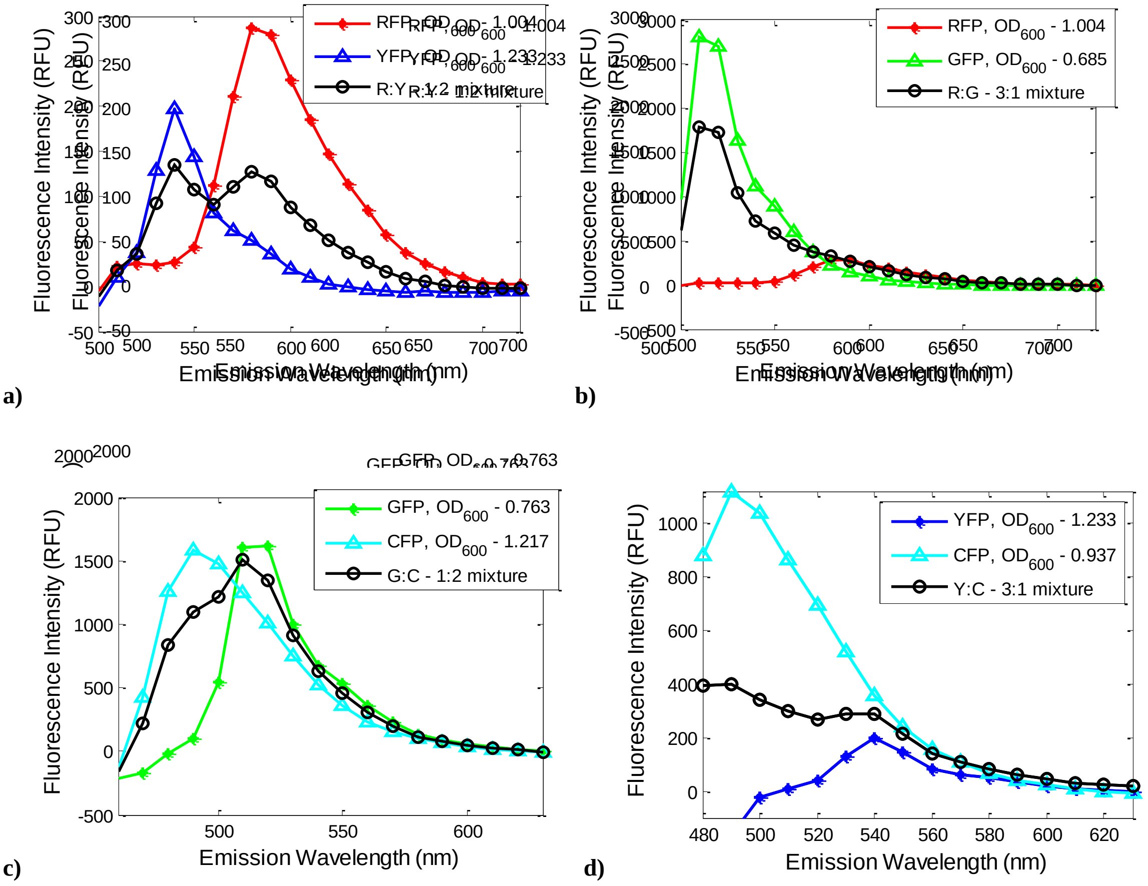


Figure 11 Emission spectra of individual proteins and their mixtures a) 1:2 mixture of red and yellow FP, excitation wavelength – 450 nm b) 3:1 mixture of red and green FP, excitation wavelength – 450 nm c) 1:2 mixture of cyan and green FP, excitation wavelength – 430 nm d) 3:1 mixture of yellow and cyan FP; excitation wavelength - 450 nm; the emission wavelengths were increased in steps of 10 nm.

Similarly, for mixtures containing three fluorescent proteins, the emission spectra of two such mixtures and the individual proteins are plotted in Figure 12. Also, a summary of the relative error in the calculated θ_j values, for all the investigated 3-protein mixtures, is given in Table 6. Again, most of the relative errors in the estimated

θ_j values are below 10%. However, the estimated θ_j values were observed to have large errors at very high O.D.s (>2.0) of the green and cyan fluorescent proteins (data not presented). This is likely caused by inaccuracies in the measurements taken by the fluorescent plate reader which were higher than the recommended intensity values.

Table 5 Mean and standard deviation (S.D.) of relative error for computing θ_j for mixtures of two proteins

Mixture		**Mean (%)**	**S.D.**
Cyan-Green	C	7.23	8.91
	G	8.69	4.91
Cyan-Yellow	C	3.43	2.07
	Y	11.55	10.64
Cyan-Red	C	14.73	5.97
	R	5.50	2.42
Green-Yellow	G	6.65	1.91
	Y	4.00	1.41
Green-Red	G	4.45	2.85
	R	7.80	6.68
Yellow-Red	Y	1.81	1.03
	R	2.00	0.85

There can also be various sources of nonlinearity in the system; for instance the overall intensity might not be a linear superposition of the intensities of individual proteins if the emission intensity of a fluorescent reporter is absorbed by another. These effects can be escalated at higher concentrations/O.D.s of the reporters in the mixtures. However, for appropriate operating conditions, the presented results validate that the overall intensity of a mixture containing multiple fluorescent proteins can be approximated as a linear superposition of the mixtures of individual proteins in the mixture.

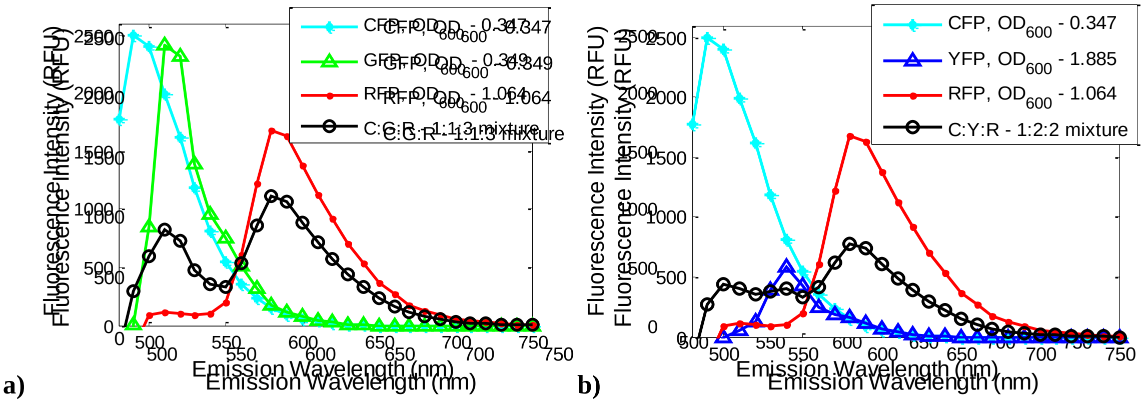


Figure 12 Emission spectra of individual proteins and their mixtures a) 1:1:3 mixture of CFP, GFP and RFP b) 1:2:2 mixture of CFP, YFP and RFP; excitation wavelength – 450 nm for both cases; the emission wavelength measurements were increased in steps of 10 nm.

4.3.2 Application of the Experimental Design Criterion

Computing the contribution of individual fluorescent proteins to the overall fluorescence intensity, as illustrated in the previous subsection, is only the first step required for model based experimental design. The second step is to determine a set of proteins to be used together such that the accuracy of estimated contributions is maximized. This subsection presents results from applying the experimental design formulation (section 4.2.5) to select the set of proteins that should be used together in an experiment.

Table 6 Mean and standard deviation (S.D.) of relative error for computing θ_j for different mixtures of three proteins

Mixture		**Mean (%)**	**S.D.**
Cyan-Green-Red	C	11.31	5.35
	G	9.96	5.15
	R	6.94	5.08
Cyan-Red-Yellow	C	9.67	4.36
	R	7.32	4.79
	Y	9.09	8.18
Green-Red-Yellow	G	5.66	4.59
	R	14.30	11.31
	Y	7.42	6.28
Cyan-Green-Yellow	C	9.73	3.46
	G	8.75	6.53
	Y	18.15	6.07

The formulation is first applied to the emission spectra of the four fluorescent proteins shown in Table 4. For this purpose, the emission spectra of these proteins, obtained at the excitation wavelength of 450nm, were normalized using their respective maximum intensities. The reason for using the normalized emission spectra for experimental design is so that the formulation is not biased against proteins with lower emission intensities caused by the choice of an excitation wavelength at which they have lower excitations. Thus, the subset of reporters is chosen solely on the basis of minimal spectral overlap and not the excitation wavelength used. These normalized spectra for CFP, GFP, YFP and RFP are shown in Figure 13.

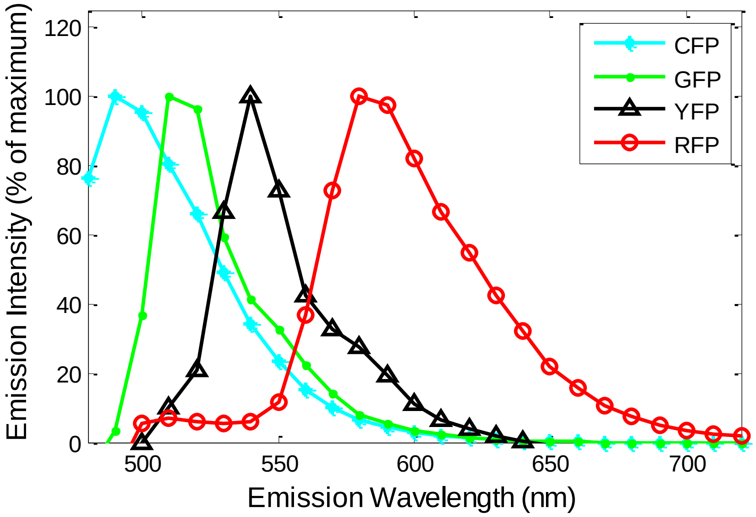
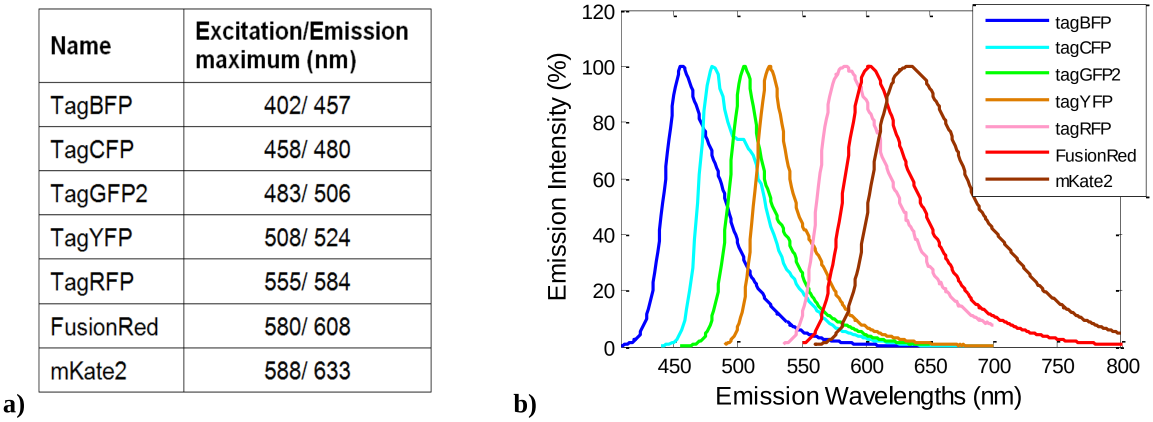


Figure 13 Normalized spectra of fluorescent proteins

On using exhaustive search, the algorithm selects the following proteins –CFP, YFP and RFP, for the selection of 3 proteins from the 4 available proteins. On close observation, it is evident from Figure 13 that the emission spectra of the chosen proteins have only minimal overlap. Furthermore, the relative errors obtained for this particular 3-protein mixture. i.e., the one containing cyan, yellow and red fluorescent protein expressing *E. coli* strains, are the lowest among the combinations that were investigated in this study (Table 6).

To further investigate the utility of the experimental design criterion for selection of fluorescent proteins from a larger set, emission spectra of 7 different fluorescent proteins were obtained from an online database (Evrogen, 2002). The details of the emissions characteristic of these proteins are summarized in Figure 14. It is evident that there is significant overlap in the emission spectra of these fluorescent proteins (Figure 14b) and it is non-trivial to choose a subset of these proteins for an experiment. For the purpose of applying the developed experimental design criterion, the emission intensities of these proteins were sampled at wavelengths ranging from 410 – 800 nm in increments of 5nm. If 4 fluorescent proteins are chosen among the available 7 proteins using the D-optimal design criterion, the result obtained from exhaustive enumeration (by evaluating $C_4^7 = 35$ protein combinations) and also from the forward selection procedure (in order of selection) is – mKate2, tagBFP, tagRFP and tagYFP. This matches with the recommendations given in (Shaner et al., 2005), based on the available filter sets, to use cyan, yellow, orange and far-red fluorescent proteins together for multiple labeling applications.



a)

Name	Excitation/Emission maximum (nm)
TagBFP	402/ 457
TagCFP	458/ 480
TagGFP2	483/ 506
TagYFP	508/ 524
TagRFP	555/ 584
FusionRed	580/ 608
mKate2	588/ 633

b)

Figure 14 a) Maximum emission and excitation wavelengths of the fluorescent proteins b) Normalized emission spectra of the fluorescent proteins

To further validate that the fluorescent protein set chosen by the D-optimal design criterion results in maximum accuracy for the estimated contributions of the individual proteins to the overall intensities, a Monte Carlo simulation study was performed to evaluate the performance of all possible protein combinations. A large number of data sets (*N*=500) were simulated for the measured overall fluorescence intensity for all 35 protein combinations. It was assumed that the actual $\Theta = [0.25\ 0.25\ 0.25\ 0.25]$ and random Gaussian noise $\sim N(0,5)$ is present in all the measurements. The overall error in the estimated Θ values for the i^{th} fluorescent protein combination was then evaluated as,

$$\text{Overall Error}_i = \mathrm{E}_{j\in\{1,\ldots,N\}}\left[\left\{(\hat{\Theta}^i_j - \Theta)^T(\hat{\Theta}^i_j - \Theta)\right\}^{\frac{1}{2}}\right] \tag{4.15}$$

where $\hat{\Theta}_j^i$ are the estimated values from the j^{th} simulated data set for the i^{th} protein combination. The expectation $\mathrm{E}[\cdot]$ is evaluated over all the simulated data sets $j \in [1,\ldots,N]$. The results obtained are plotted in Figure 15.

In the figure (Figure 15), the lowest overall error is obtained for the 18th subset which contains mKate2, tagBFP, tagRFP and tagYFP reporters which is also the same combination selected by using the D-optimal design criterion.

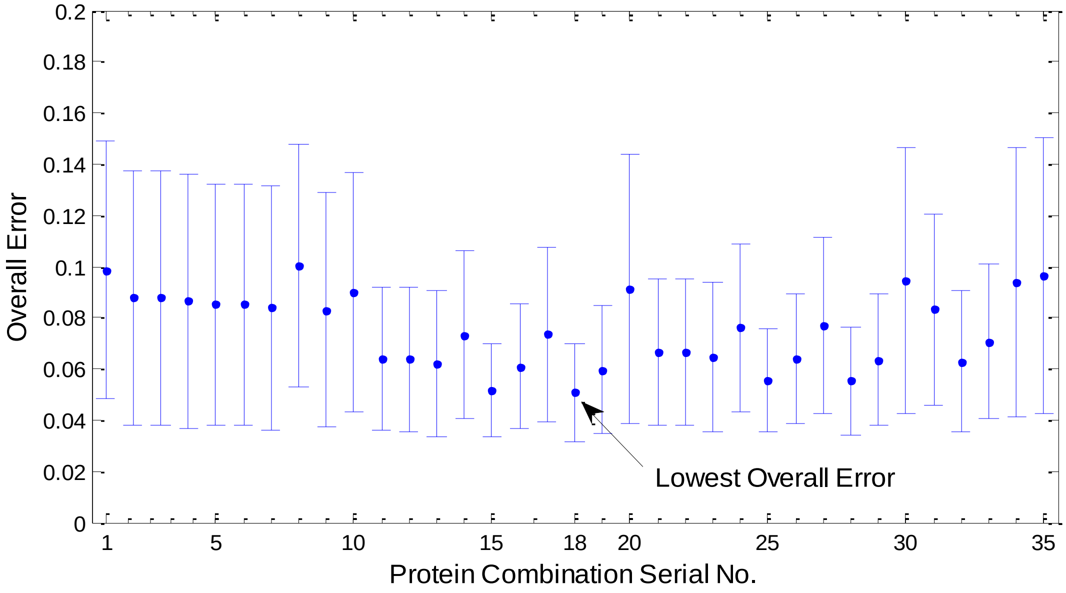


Figure 15 Overall error with ± standard deviation in estimated Θ values for all the protein combinations by choosing 4 out of 7 commercially available fluorescent proteins.

4.4 Discussion and Summary

An experimental design criterion for selection of fluorescent proteins that can be simultaneously used in an experiment has been developed. It is the goal of this experimental design to maximize the accuracy at which the contribution of individual proteins can be extracted from the overall intensity of the sample. For this purpose the overall fluorescence intensity of the mixture measured using a plate reader is represented by a linear superposition of the intensities of the individual proteins. Using this formulation, the average of the relative errors for computing the contributions of individual proteins in mixtures containing 2 and 3 fluorescent proteins were obtained to be 6.41± 3.96 % and 9.86± 3.45%, respectively (from data in Table 5 and Table 6). These errors are within the acceptable limits for biological data and hence verify that the linear unmixing formulation (Dickinson et al., 2001) which has previously only been applied to pixels in images from fluorescence microscopy can also be applied to emission spectra of mixtures measured using a plate reader. While there are a few cases in which the relative error obtained is larger than the aforementioned values, the results are still within acceptable limits. These larger errors can be due to factors such as insufficient emission intensity of red and yellow fluorescent proteins at the chosen excitation wavelength, inaccurate readings by the plate reader at very high emission intensities of proteins such as GFP and CFP and absorbance of emission intensity of one reporter by the other.

In a second step, the D-optimal design criterion is used for model-based design for optimal selection of fluorescent proteins for simultaneous use. The resulting mixed integer programming problem is solved using exhaustive search and, as a sub-optimal, but computationally inexpensive alternative, a forward selection algorithm. It has been successfully applied to the emission spectra of four fluorescent proteins used in this work and another set of seven fluorescent proteins for which the emission spectra are obtained from an online database (Evrogen, 2002). Thus, the developed design criterion can be used for screening of proteins for applications where multiple events need to be monitored and it can be used in addition to other considerations which are currently used, such as stability, toxicity, and maturation time of fluorescent proteins.

5. MODELING OF CELL POPULATIONS LABELED WITH A FLUORESCENT REPORTER SYSTEM

5.1 Introduction

Fluorescent proteins have been widely used as markers of gene expression and transcriptional regulation due to several reasons (Chalfie et al., 1994), such as their plasmid can be easily integrated into the DNA of the cells, they do not have a toxic effect on cell growth and they can also be conveniently detected by illuminating the sample with suitable light. Thus, as discussed in Section 3, the fluorescence obtained in fluorescent protein based reporter systems can be used as an indicator of transcription and translation. Furthermore, in section 3, an inverse problem has been formulated and solved to estimate the overall dynamics of transcription factors using the average dynamic fluorescence intensity data from a GFP reporter system. However, there are phenotypic variations in cell populations due to which each cell exhibits different fluorescence intensities. The fluorescence observed in reporter systems can be affected by noise in gene expression (Swain et al., 2002) as well as physiological factors such as unequal partitioning of cellular material resulting from cell division (Hjortsø, 2006). Since most of the experimental data of fluorescence is obtained from techniques such as flow cytometry or fluorescence microscopy that utilize cell populations, using single cell models to estimate cell physiological parameters or transcriptional dynamics may lead to erroneous conclusions (Hasenauer et al., 2011a). Furthermore, it is non-trivial to track individual cells in a population during the course of an experiment to analyze how fluorescence in a single cell evolves over time (Huang et al., 2012). Thus, in this section, a dynamic model for a cell population that is labeled with a fluorescent protein reporter system has been developed to describe the dynamics of the fluorescence intensity distribution of the cells.

Several examples exist in literature where population balance equation (PBE) modeling (reviewed in section 2.4.1) has been used for describing the dynamics of cells populations by using structuring variables such as cell mass (Mantzaris et al., 1999; Mhaskar et al., 2002; Zhu et al., 2000), intensity of fluorescent dyes (Banks et al., 2010; Luzyanina et al., 2007), number of plasmids (Ganusov et al., 2000) or cell age (Gabriel et al., 2012). In this work, the PBE modeling technique has been adopted for modeling fluorescent protein-labeled cell populations by using the fluorescence intensity of the reporter protein as the structuring or independent variable. The main reason for using PBE modeling technique is that it provides a better framework for estimating the unknown transcriptional or physiological parameters for cell populations (Abu-Absi et al., 2003; Friedrich, 1999; Hjortsø, 2006) as compared to the cell ensemble modeling technique (section 2.4.2). Furthermore, there are few examples in literature (Banks et al., 2011a; Banks et al., 2011b; Luzyanina et al., 2007) where the developed population balance model (PBM) has been validated using experimental data and the unknown parameters describing the single cell physiological functions have been estimated. In this work, we have attempted to validate the developed PBE model for a GFP reporter cell line of HeLa cells, containing the Tet-on expression system, which has previously been developed by our collaborators (Huang et al., 2010b). Thus, a comprehensive study is done involving cell population balance model development, its solution as well as model validation using experimental data for a cell population containing fluorescent protein based reporter system.

The following subsections describe the details of the model development scheme for fluorescent cell populations (section 5.2), the solution of the dynamic model using a finite difference method (section 5.3), and model validation using flow cytometry data of HeLa cells containing the Tet-on expression system (section 5.4). Then, the limitations of the developed approach are discussed followed by a brief summary.

5.2 Model Development

As mentioned in section 2.4.1, cell population balance modeling requires a dynamic balance over the various factors that affect the independent variable of interest, which in our case is the fluorescence intensity of cells. There are a few examples in literature, for instance (Luzyanina et al., 2007), where the PBMs have been developed based on fluorescence intensity. However, these models are based on cell populations which are labeled with fluorescent dyes and the dye is not continuously produced inside the cell. For cell populations in which a fluorescent protein is produced as a result of gene expression, the factors affecting the resulting fluorescent intensity distribution can be summarized as,

1) There is an increase in fluorescence intensity in single cells due to the production of the fluorescent protein due to gene expression.
2) There is a decrease in fluorescence intensity due to cell death.

3) When cells divide, the fluorescent protein, which is present in the cytoplasm of the cells, is divided between the two daughter cells. In this work, this division is assumed to be conservative i.e. the entire fluorescent protein from the mother cell is divided among daughter cells.

4) Furthermore, when the cell divides the distribution of the fluorescent protein, which is present in the cytoplasm, is assumed to be unequal between the two daughter cells. The distribution is described by a partition function.

Taking into account these factors, the PBM is written as,

$$\begin{aligned}&\frac{\partial W(z,t)}{\partial t}+\frac{\partial\big(r(z)W(z,t)\big)}{\partial z}\\&= -\big(\Gamma(z)+\Theta(z)\big)W(z,t)+\ 2\int_0^\infty \Gamma(z')p(z,z')W(z',t)dz'\end{aligned} \tag{5.1}$$

This is a first order integro-partial differential equation with fluorescence intensity z and time t as independent variables and the number distribution of cells - $W(z,t)$ as the dependent variable. In the above equation, $r(z)$ is the net growth rate of fluorescence intensity, $\Gamma(z)$ and $\Theta(z)$ are the division and death rate respectively and $p(z,z')$ represents the partition function. The integral term $2\int_0^\infty \Gamma(z')p(z,z')W(z',t)dz'$ gives the rate at which the new cells are formed with fluorescent intensity z and it is obtained by integrating over the entire range of the fluorescence intensities of the dividing cells. The physiological functions i.e. $r(z)$, $\Gamma(z)$, $\Theta(z)$ and $p(z,z')$ are unknown for any cell population. These functions describe the behavior of single cells and are crucial part of forming a PBM. Mostly, there is little information available about these functions, however, they can be chosen by making reasonable assumptions or prior knowledge about the system of interest. The choices for these functions for the fluorescence intensity structured PBM developed in this work are given below.

1) The net growth rate of fluorescence intensity (FI) is assumed to be linear w.r.t. z such that

$$r(z) = kz. \tag{5.2}$$

It takes into account the factors contributing to FI growth as well as its decrease due to degradation of fluorescent proteins.

2) The division rate is assumed to be a Gaussian function of FI as the division rate has been observed to be bell shaped w.r.t. intensity in some studies (Luzyanina et al., 2007).

$$(z) = a \exp\left(-\frac{(z-u_d)^2}{2{s_d}^2}\right) \tag{5.3}$$

3) The death rate can be assumed to be a constant,

$$\Theta(z) = dr \tag{5.4}$$

4) A very commonly used form for the partition function, $p(z, z')$ is a symmetric beta distribution function (Mantzaris et al., 1999; Nikos V, 2006) such that,

$$p(z, z') = \frac{1}{B(q,q)}\left(\frac{z}{z'}\right)^{q-1}\left(1-\frac{z}{z'}\right)^{q-1}\frac{1}{z'} \tag{5.5}$$

where q is a parameter of the distribution and $B(q,q)$ is the beta function. $p(z,z')$ gives the fraction of the newborn cells formed with fluorescence intensity z and $z'-z$ when a cell with intensity z' divides.

Thus, the cells are not assumed to be dividing equally and a distribution of the partition of the fluorescent protein or its intensity is obtained when cells divide and it is described by the partition function presented above.

5.3 Model Simulation Using a Finite Difference Scheme

This subsection discusses the implementation of the finite difference scheme for the population balance model developed in the previous section. This is followed by an illustration of the results obtained from simulating the model for nominal values of the parameters in the PBM.

5.3.1 Finite Difference Scheme

The developed model (equation (5.1)-(5.5)) is an integro-partial differential equation (IPDE) and its initial boundary value problem (IBVP) is numerically solved using an implicit finite difference scheme. Particularly, the Crank Nicolson scheme (Smith, 1985) is used as it resulted in stable solutions for the presented partial differential equation. The scheme is second order accurate in both the independent variables i.e. time as well as fluorescence intensity. The steps for implementation of the Crank Nicolson scheme are discussed next.

The boundary conditions represent that there are no cells beyond the two extremes of the distribution and they are given by the following equations.

$$W(z_0, t) = 0 \qquad (5.6)$$

$$W(z_m, t) = 0$$

where z_m is chosen few units larger than the maximum fluorescence intensity observed in the system and z_0= 0 such that the above boundary condition is satisfied. Furthermore, the initial distribution of FI in the cell population is assumed to be known, thus the initial condition is,

$$W(z, t_0) = w_0(z) \qquad (5.7)$$

where t_0 is the initial time point.

Let the time steps and the grid size for FI be denoted by Δt and Δz respectively, such that,

$$z = z_0 + i\,\Delta z \;\; \forall\, i = \{0,1,\ldots,\mathrm{M}\} \qquad (5.8)$$

$$t = t_0 + j\,\Delta t \;\; \forall\, j = \{0,1,\ldots,\mathrm{N}\}$$

In Crank Nicolson scheme, the partial time derivatives are approximated using a forward finite difference,

$$\frac{\partial W(z,t)}{\partial t} = \frac{W_i^{j+1} - W_i^{j}}{\Delta t} \qquad (5.9)$$

where $W_i^j = W(z_0 + i\Delta z, t_0 + j\Delta t)$. The partial z space derivatives are approximated as an average of the centered space difference at current (j^{th}) and the next ($j+1^{th}$) time step i.e.

$$\frac{\partial r(z,t)W(z,t)}{\partial z}=\frac{1}{2}\left[\left(\frac{r_{i+1}W_{i+1}^{j+1}-r_{i-1}W_{i-1}^{j+1}}{2\Delta z}\right)+\left(\frac{r_{i+1}W_{i+1}^{j}-r_{i-1}W_{i-1}^{j}}{2\Delta z}\right)\right] \tag{5.10}$$

The integral term in equation (5.1) is approximated using the trapezoidal rule for numerical integration. For the purpose of numerical integration, the integral term $2\int_0^\infty \Gamma(z')p(z,z')W(z',t)dz'$ can be written as $2\int_z^{z_m} \Gamma(z')p(z,z')W(z',t)dz'$. This is because to evaluate the rate at which cells with fluorescence intensity z are being produced, the diving cells are likely to have intensities $z' \in [z, z_m]$ where $z_m = z_0 + \text{M}\Delta z$ is the upper limit for the fluorescence intensity observed in the system. Thus the finite difference equation for the presented IPDE can be formulated as,

$$\begin{aligned}\frac{W_i^{j+1}-W_i^j}{\Delta t}+\frac{1}{2}\left[\left(\frac{r_{i+1}W_{i+1}^{j+1}-r_{i-1}W_{i-1}^{j+1}}{2\Delta z}\right)+\left(\frac{r_{i+1}W_{i+1}^{j}-r_{i-1}W_{i-1}^{j}}{2\Delta z}\right)\right]\\ =-(\Gamma_j+\Theta_j)W_i^j+2\Delta z\left(\sum_{i'=i+1}^{M-1}\Gamma_{i'}p_{i,i'}W_{i'}^j+\frac{\Gamma_M p_{i,M}W_M^j}{2}\right)\end{aligned} \tag{5.11}$$

For a complete derivation of the approximation of the integral term using the trapezoidal rule, please refer to Appendix B. Equation (5.11) can be reformulated as

$$\begin{aligned}\alpha_{i-1}W_{i-1}^{j+1}+W_i^{j+1}+\alpha_{i+1}W_{i+1}^{j+1}=(1-\beta_i-\Theta_i\Delta t)W_i^j-\alpha_{i+1}W_{i+1}^j+\alpha_{i-1}W_{i-1}^j\\ +2\Delta z\left(\sum_{i'=i+1}^{M-1}\beta_{i'}p_{i,i'}W_{i'}^j+\frac{\beta_M p_{i,M}W_M^j}{2}\right)\end{aligned} \tag{5.12}$$

where $\alpha_i=\dfrac{r_i\Delta t}{4\ \Delta z}$ and $\beta_i=\Gamma_i\Delta t$. This is a finite difference equation in which the values of the dependent variable at $j+1^{th}$ time point is calculated using the values at j^{th} time point such that at every time step $t = t_0 + j\Delta t$, $\forall\, j=\{1,2,\ldots,\text{N}\}$ a linear system of equations is solved to obtain the values of $W_i^{j+1}\ \forall\ i = \{1, 2, \ldots, \mathrm{M}-1\}$. The system of equations is given by

$$\begin{bmatrix} \alpha_0 & 1 & \alpha_2 & 0 & \ldots & 0 \\ 0 & \alpha_1 & 1 & \alpha_3 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \alpha_{M-2} & 1 & \alpha_M \end{bmatrix} \begin{bmatrix} W_0^{j+1} \\ W_1^{j+1} \\ \vdots \\ W_M^{j+1} \end{bmatrix} = \begin{bmatrix} (\mathrm{R.H.S})_1^j \\ (\mathrm{R.H.S})_2^j \\ \vdots \\ (\mathrm{R.H.S})_{M-1}^j \end{bmatrix} \tag{5.13}$$

where

$$\begin{aligned} (\mathrm{R.H.S})_i^j &= (1-\beta_i - \Theta_i \Delta t) W_i^j - \alpha_{i+1} W_{i+1}^j + \alpha_{i-1} W_{i-1}^j \\ &\quad + 2\,\Delta z \left(\sum_{i'=i+1}^{M-1} \beta_{i'} p_{i,i'} W_{i'}^j + \frac{\beta_M\, p_{i,M} W_M^j}{2} \right) \end{aligned} \tag{5.14}$$

All the simulations in the presented work are done using MATLAB.

5.3.2 Results for Model Simulation

This section briefly describes the results from simulating the developed cell population balance model using the finite difference scheme discussed in the previous section. Also, the effect of the various parameters, used in the physiological functions, on the resulting cell number distributions is also presented. For illustration purposes, in this section the model is simulated using the nominal values of the parameters which are chosen either based on values used in literature or on the author's experience of working with cell populations containing fluorescent reporter systems.

The initial distribution for simulating the model is assumed to be Gaussian, i.e.

$$W(z,t_0) = \frac{1}{\sigma_0\sqrt{2\pi}}\exp(-\frac{(z-\mu_0)^2}{2\sigma_0{}^2}) \tag{5.15}$$

where the values for μ_0 and σ_0 are the mean and the standard deviation of the distribution. The values used for all the parameters are given in Table 7.

Table 7 Nominal values of the parameters used for simulating the model

Parameter	Value		Parameter	Value
k	0.01 [hr^{-1}]		μ_0	850
a	0.08 [hr^{-1}]		σ_0	50
u_d	1100 [RFU]		z_0	0 [RFU]
s_d	250 [RFU]		t_0	0 [hr]
q	40		Δz	10 [RFU]
dr	0.01 [hr^{-1}]		Δt	0.01 [hr]

RFU = Relative Fluorescence Units

The model is simulated for t_n= 48 hours for $z \in [0,1500]$ and a 3D plot of the resulting dynamic cell number distribution w.r.t. fluorescence intensity and time is shown in Figure 16. It can be observed from this plot that different generations of cells are obtained as they undergo cell division with subsequent generations having lower fluorescence intensities.

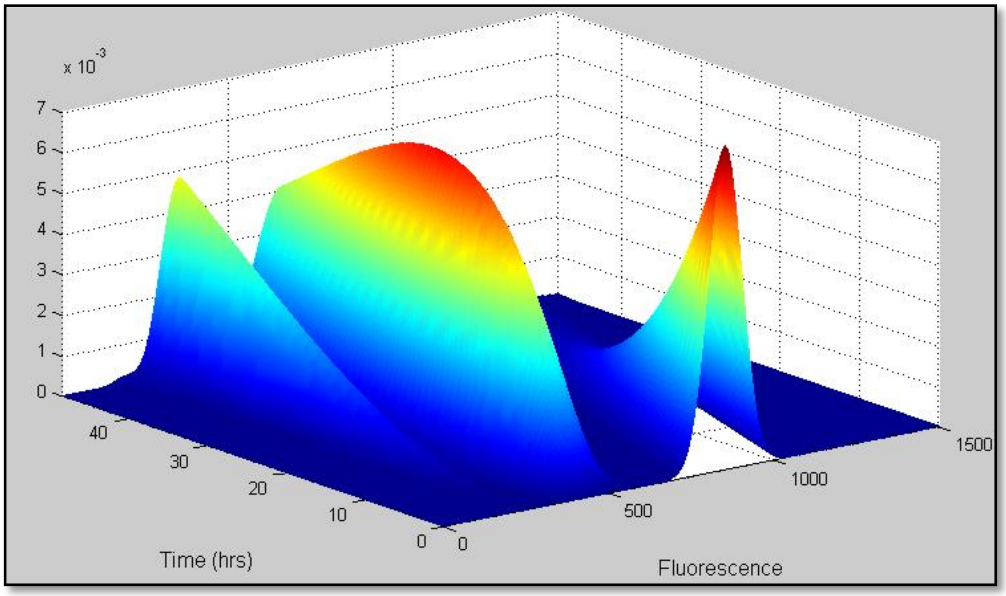


Figure 16 A 3D plot of cell number distributions w.r.t. fluorescence intensity and time, simulated using nominal values of the parameters

A 2D plot of the distributions obtained after 24 and 48 hours is also plotted along with the initial distribution in Figure 17 to have a closer comparison. While, each subsequent generation of cells have smaller fluorescence intensities, it can also be observed that each peak is also shifted towards higher fluorescence intensities as there is a net growth of fluorescence intensity in cells. Furthermore, it is non-trivial to empirically develop a model describing the dynamics of multimodal distributions like the one shown in Figure 16 and Figure 17 and requires first principle modeling to explain the complex dynamics.

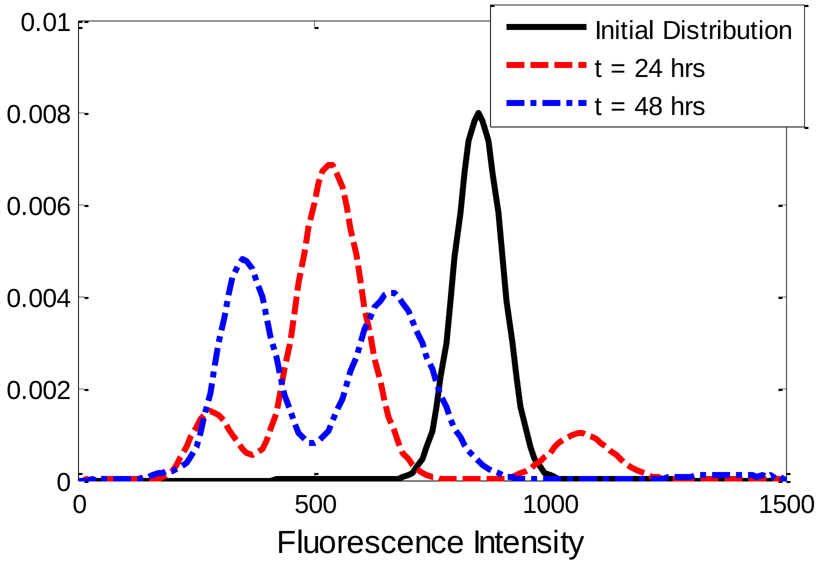


Figure 17 Cell number distributions obtained by simulating the PBM using nominal values of the parameters

Next, to analyze the affect of different parameters in the physiological functions on the resulting cell number distributions, some of the parameters are varied one by one and the resulting distributions are compared with the distribution shown in Figure 17 obtained using the nominal values in Table 7.

When the value of the net growth rate of fluorescence intensity i.e. k is doubled (Figure 18(a)), the resulting peaks are shifted towards higher fluorescence intensities. Since the division rate is also a function of the fluorescence intensity z, this causes more cells to divide earlier and the resulting cell number distribution at 24 hours is widely different from its counterpart in Figure 17 which is based on nominal values. Next, the mean of the division rate function i.e. u_d is shifted to higher fluorescence intensity of 1600 RFU as compared to 1100 RFU (Figure 18(b)). Thus, most of the cells will not divide until the FI growth causes their intensity to reach values around 1600 RFU and this causes a delay in formation of the next generation of cells. The total number of cells are also lesser in this case (Figure 19).

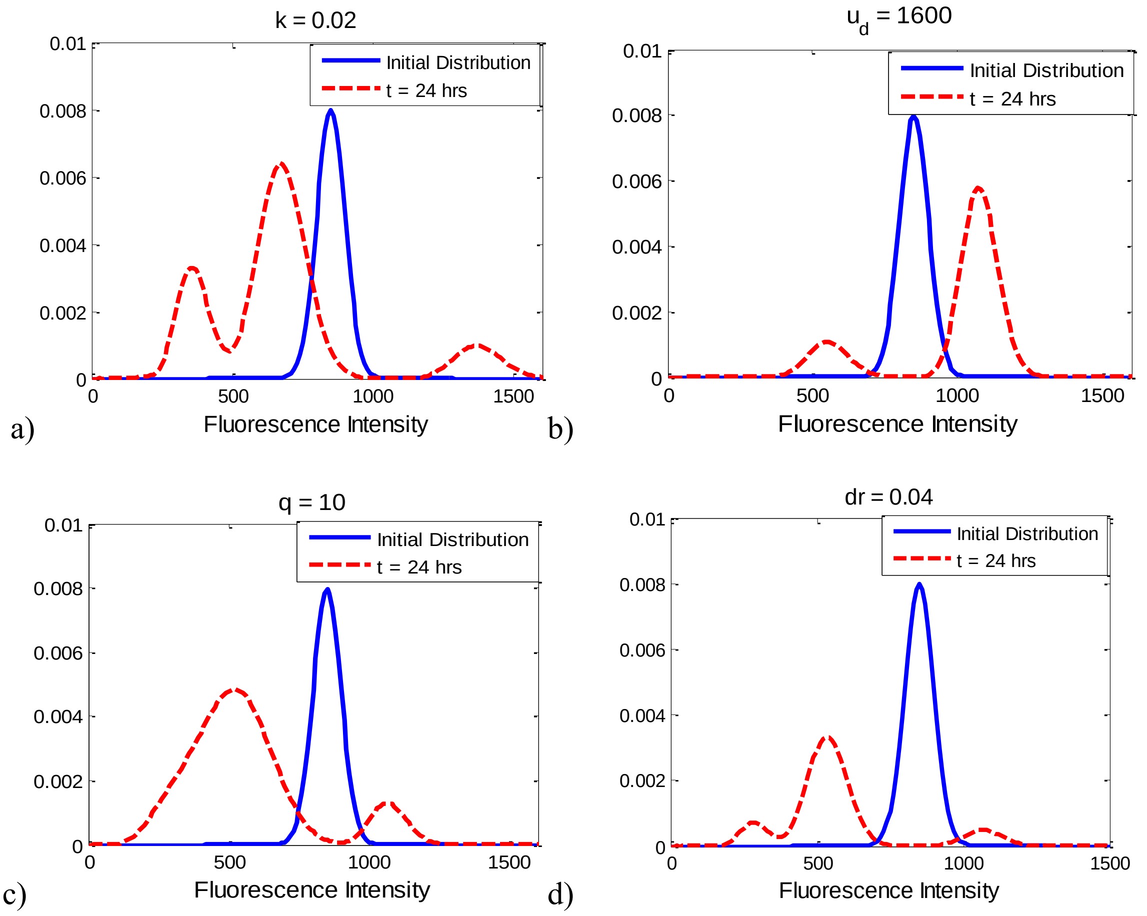


Figure 18 Simulation of the PBM using nominal values of the parameters and changing any one parameter at a time a) The parameter k is changed to 0.02 hr^{-1} b) u_d = 1600 RFU c) q = 10 d) dr = 0.04 hr^{-1}

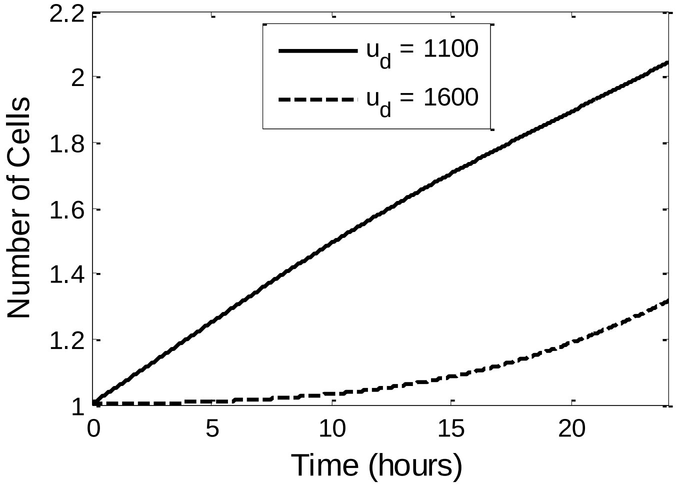


Figure 19 Comparison of the total number of cells in the system when the value of the parameter u_d is changed from 1100 RFU to 1600 RFU

Figure 18(c) shows the cell number distributions when the value of the parameter of the partition function i.e. q is decreased from its nominal value. A smaller value of q results in formation of new cells over a wide of range of fluorescence intensities (FI) as compared to when cells divide with almost equal division of FI among the daughter cells. Thus, it is now difficult to differentiate between the cell generations or the different peaks in the distribution. The last subfigure (Figure 18(d)) illustrates that at a higher cell death rate, the cell number distribution has lower values because of a decrease in the number of viable cells. Also, the decrease is observed over the entire range of FI since the death rate is a constant and independent of the FI.

In the next subsection, the unknown parameters of the developed cell population balance model are estimated using experimental data for a fluorescent reporter system of HeLa cells.

5.4 Validation of the Developed Model Using Flow Cytometry Data

This section discusses the procedure for estimation of the unknown parameters in the cell population balance model. The first subsection describes how the experimental data was obtained by our collaborators. Then, before fitting the data to the PBM, a global sensitivity analysis is being carried out on the PBM to identify the parameters that can be easily estimated using the experimental data. Then, the 'identifiable' parameters are estimated using experimental data.

5.4.1 Obtaining Experimental Data

The experimental data was obtained for a previously developed cell line of HeLa cells containing the Tet-on expression system (Huang et al., 2010b). This system consists of an artificial inducible transcription factor tTA which is activated by addition of doxycycline (Dox). The plasmid for GFP is cloned on the response plasmid downstream of the response element. Then, the GFP expression is initiated by tTA-Dox complex when it translocates to the nucleus and binds to the RE. The reason for using the "HeLa Tet-on" system in this study is that firstly the P expression is easily inducible by addition of Dox and there is no complex signal transduction dynamics involved. And secondly, because HeLa cells have a moderately faster growth rate as compared to other mammalian cell lines. Then, the fluorescence intensity distributions of the cells were obtained using flow cytometry. A brief overview of the experimental procedure performed by Ms. Shreya Maiti in Dr. Arul Jayaraman's laboratory (Chemical Engineering, Texas A&M University) is described below.

HeLa cells containing the Tet-on expression system along with the GFP reporter plasmid were seeded on a 96 well plate and incubated at 37ºC for 4 hours to let them attach to the surface. They were then stimulated with two different concentrations of Dox (1 μg/ml and 10 μg/ml). The cells were monitored for a total of 4 days i.e. 96 hours with 1/3 of the medium (with corresponding volumes of Dox) replenished in the wells every 24 hours. This is done so that the cells are not deprived of the nutrients essential for cell growth and division. The cells were harvested 2, 12, 24, 36, 48, 60, 72, 84 and 96 hours after stimulation and fixed with 4% paraformaldehyde to prevent the degradation of GFP present in the cells. The fixed cells were analyzed using flow cytometry in which 10,000 cells were sampled from every well and their fluorescence intensities were measured to obtain the fluorescence intensity distributions. The flow cytometer BD FACSAria II was used for this purpose.

The software FlowJo, commonly used for analyzing flow cytometry data, was used for performing gating on the samples to separate the cell debris from the measurements of live cells in the data set. This was done by excluding the measurements corresponding to particles which have a small value of forward scattered (FSC) as well as side scattered light (SSC) (Biosciences, 2000). FSC is a measure of cell volume and SSC is related to the inner complexity and integrity of the cells. Thus, particles with smaller values of FSC as well as SSC are probably dead cells or other debris and should be excluded from the data set (Shapiro, 2005). The resulting FI distributions for HeLa cells stimulated with 1 μg/ml and 10 μg/ml of Dox are plotted in Figure 20.

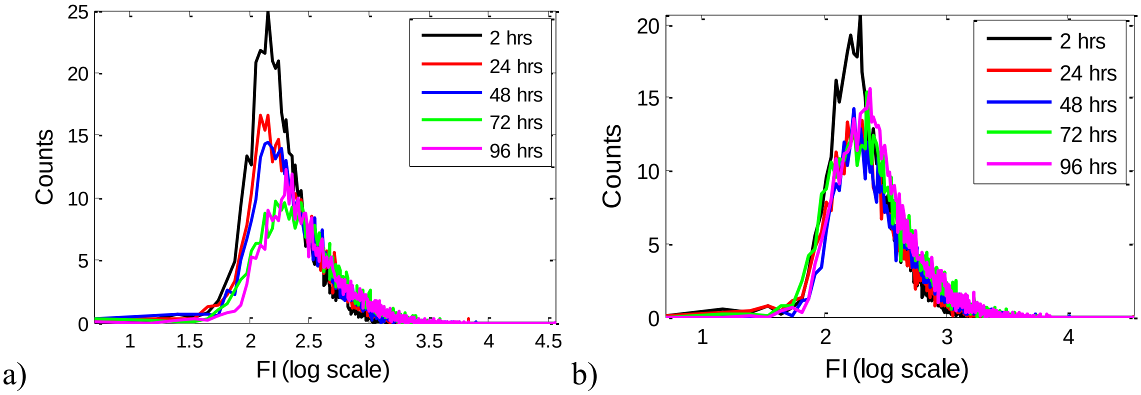


Figure 20 Fluorescence intensity (FI) distributions obtained for HeLa cells using flow cytometry when they are stimulated with a) 1 µg/ml of Dox 2) 10 µg/ml of Dox

These distributions were smoothed using the moving average filter and the resulting distribution for the data set stimulated with 1 µg/ml of Dox is shown in Figure 21. The smoothed distributions were used for model validation.

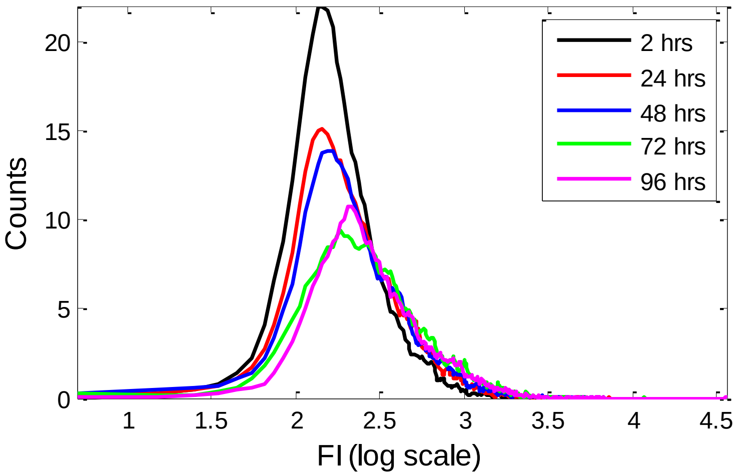


Figure 21 Smoothed fluorescence intensity distributions obtained using the moving average filter for cells stimulated with 1µg/ml of Dox

5.4.2 Global Sensitivity Analysis of the PBM

There are 6 unknown parameters in the formulated PBM given by the net growth rate of fluorescence intensity k, the parameters describing the division rate i.e. a, u_d and s_d, the partition function parameter q and the death rate dr. However, it may not be possible to estimate all the parameters using the measurements of fluorescence intensity cell number distributions. This is because some of these parameters might not have significant effect on the FI distributions. Also, the parameters can even have correlated effects on the output such that the change in one value of the parameter can be compensated by a change in the value of another parameter. This can cause the optimization algorithms to struggle in obtaining an optimum value for all the parameters. Thus, a global sensitivity analysis for the developed population balance model with respect to the unknown parameters is performed. This is done to identify the parameters that can be easily estimated using data for FI distributions and also to see if any of the parameters have correlated effects on the output modeled distributions.

Morris method (reviewed in section 2.5) was used for calculating the sensitivities of the output distributions w.r.t. all the parameters. Furthermore, the sensitivities are calculated for the normalized distributions from the cell population balance model instead of the cell number distributions. This is because only a fixed number of cells are sampled from each well in flow cytometry for measurement of fluorescence intensities and thus the total number of cells are not known at each time instant. The uncertainty range of the parameters is chosen to be ± 100% around the nominal values used for model simulation in section 5.3.2 and the initial distribution is also assumed to be Gaussian.

The results for the mean and standard deviation of the elementary effects for all the parameters obtained using Morris method are shown in Figure 22. The statistics for the elementary effect of death rate is not shown in this figure because both the mean and standard deviation is of the order of 1e-10 for death rate. This implies that the death rate has negligible effect on the resulting normalized FI distributions. It can also be observed from the model simulations given in Figure 18 in the previous section where a variation in the death rate only changes the intensity of the resulting distribution but the shape of the distribution remains the same.

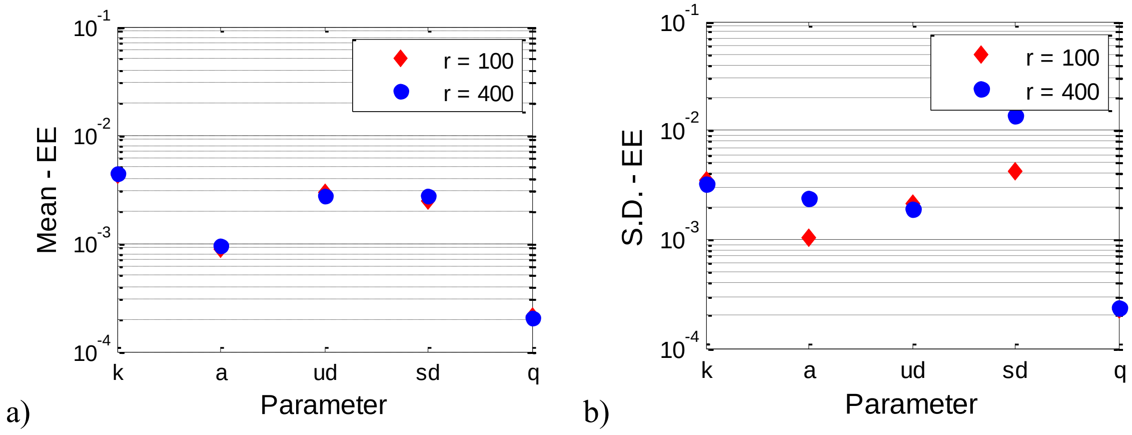


Figure 22 Statistics of the elementary effects (EE) obtained using Morris method where r denotes the number of EEs sampled a) mean of the EE b) standard deviation (S.D.) of the EE

It can be observed that the parameter q does not have a significant effect on the output and k, u_d and s_d seem to be most important input parameters. However, a larger value of the standard deviation of the elementary effect of s_d also implies that it is involved in correlated effects with other parameters. Thus, only k and u_d were considered for parameter estimation.

5.4.3 Estimation of Unknown Parameters: Preliminary Results and Discussion

After performing the sensitivity analysis w.r.t. to the unknown parameters in the cell PBM, k and u_d were recognized to be the identifiable parameters while the other parameters i.e. a, s_d, q and dr were kept constant. Furthermore, the death rate dr is assumed to be negligible since the cells are replenished with 1/3rd of the fresh medium every 24 hours to reduce cell death.

The function 'fmincon' in MATLAB was used for parameter estimation by minimizing the least squares fitting error between the normalized FI distributions obtained from flow cytometry and from the simulation of the PBM using the finite difference scheme. The FI distribution measured at 2 hours after stimulation with Dox for HeLa cells was used as an initial distribution for simulating the model to allow time for induction of GFP. The values of time step and grid size used are $\Delta t = 0.01$ hours and Δz = 10 RFU. The model is simulated for $t_n = 96$ hours and for FI ∈ [0, ~3.2e+3 RFU]. With these specifications, each simulation of the model took approximately 5-6 seconds for execution on an Intel Core i5 CPU (2.53 GHz). To decrease the computation time during parameter estimation, where several objective function evaluations are typically required for calculation of the gradients in optimization algorithms, the parallel computation capabilities of MATLAB (Sharma and Martin, 2009) were utilized such that the function evaluations can be executed simultaneously on multiple cores of the processor.

Since the values of the unknown parameters i.e. a, s_d and q cannot be efficiently estimated using optimization algorithms, their reasonable values were chosen using the following study. A large number of values of these 3 parameters were sampled using latin hypercube sampling (LHS) and the optimization problem for estimation of k and u_d using the experimental data was solved for each of these samples. The value of the parameter set which resulted in lowest least squares fitting error between the simulated and experimental data for HeLa cells, was chosen as the fixed values for the parameters to be kept constant. The normalized values of the least squares fitting error for experimental data of HeLa cells stimulated with 1 μg/ml of doxycycline is plotted against the sampled values of a and s_d in Figure 23. It can be observed that multiple minima's are obtained in the fitting error for various values of the parameter a while all other parameters are varied. This implies that the parameter a is not useful in describing any significant trends in this data and it may be obsolete for this particular data set. Similar plots and observations were obtained for the parameter q as well for both the data sets in which HeLa cells were stimulated with 1 μg/ml and 10 μg/ml of doxycycline. However, the parameter s_d seems to be significant in describing the variation in the data and there is a particular value of s_d for which the minima is obtained.

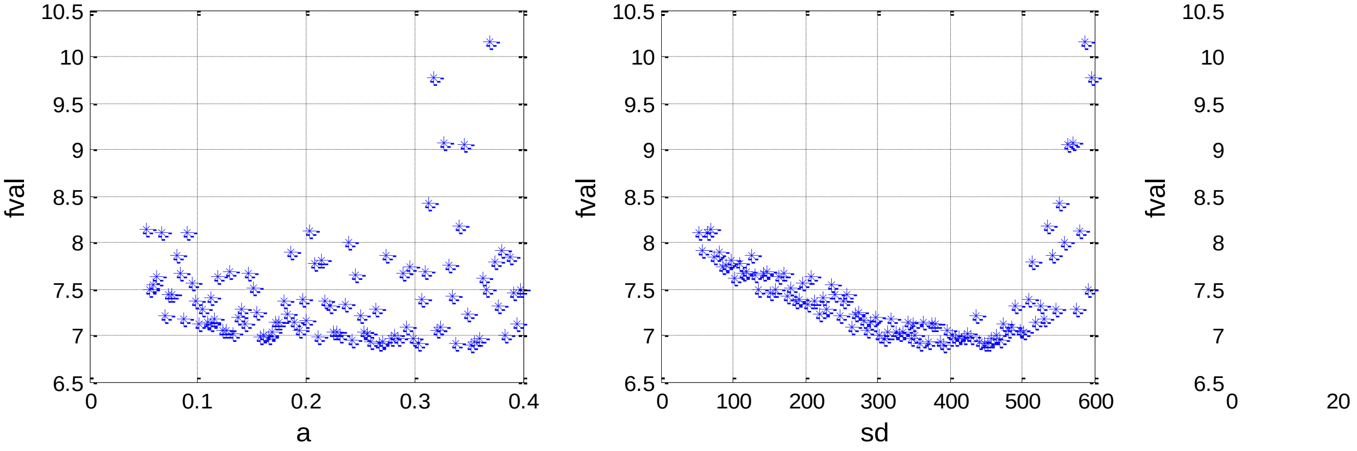


Figure 23 Plots of normalized fitting error vs two parameters a [hr^{-1}] and s_d [RFU] from the solution of the parameter estimation problem for different parameter sets sampled using LHS

After selecting the values of a, s_d and q, the net growth rate of FI i.e. k and the mean of the division rate physiological function u_d were estimated using experimental data by using several initial values for optimization. The initial values were again sampled using latin hypercube sampling and the parameter values that resulted in lowest fitting error were selected as the estimated values. Furthermore, the optimization problem was solved for different subsets of the available data by leaving the data for one time point at a time. The mean and standard deviation of the two estimated parameters calculated by solving the optimization problem from different subsets of data along with the parameters that were kept constant are given in Table 8.

Table 8 Parameters estimated using experimental data for HeLa cells stimulated with 1 μg/ml and 10 μg/ml of Dox

Parameter [units]	1 μg/ml Dox stimulation	10 μg/ml Dox stimulation
Net growth rate of FI (k [hr^{-1}])	0.0063 ± 0.00031	0.0041 ± 0.00026
Division Rate function Mean (u_d [RFU])	2179.74 ± 24.88	2452.23 ± 22.07
Division Rate Intensity a ([hr^{-1}])	0.3523	0.3969
Division Rate function S.D. (s_d [RFU])	391.7	437.25
Partition Function parameter (q)	55.4	36.4

RFU = Relative Fluorescence Units

The Tet-on expression system is expected to produce increased expression of the reporter protein on increasing the concentrations of Dox. However, from the estimated values of the parameters (Table 8), it can be observed that the net growth rate of fluorescence intensity is observed to be lesser for HeLa cells stimulated with 10 μg/ml of Dox than for cells stimulated with 1 μg/ml of Dox. This is probably because when cells are being monitored for extended periods of time, like in this presented study, higher concentrations of Dox can have a detrimental effect on cancer cell lines (Fife and Sledge Jr, 1998; Mouratidis et al., 2007; Onoda et al., 2006). This can affect the transcription rate of the Tet-on expression system and may even cause cell death. Thus, the net growth rate of intensity for the Tet-on expression system in cancer cell lines, averaged over a period of 3-4 days, is likely be lesser for higher stimulating concentrations of Dox.

Furthermore, it can also be observed from the estimated values of u_d that the division rate Gaussian function is shifted towards higher fluorescence intensities on increasing the concentrations of doxycycline. From the FI intensity distributions (Figure 20) it can be seen that a large portion of the cells have fluorescence intensities less than 1e+3 RFU. This implies that the division rate of cells may have decreased with an increase in doxycycline concentration. This can also be attributed to the negative effect of Dox on cell proliferation and growth in cancer lines (Fife and Sledge Jr, 1998; Onoda et al., 2006).

The plots of the fits between the normalized distributions of FI from experimental data and model simulations are given in Figure 24 for HeLa cells stimulated with 1 μg/ml doxycycline. The model simulation is carried out using the parameter values given in Table 8. It can be seen that the plots observed are satisfactory, however there is considerable scope for improvement in the theoretical population balance model as well as in using more reliable ways of carrying out the experiments and obtaining experimental data.

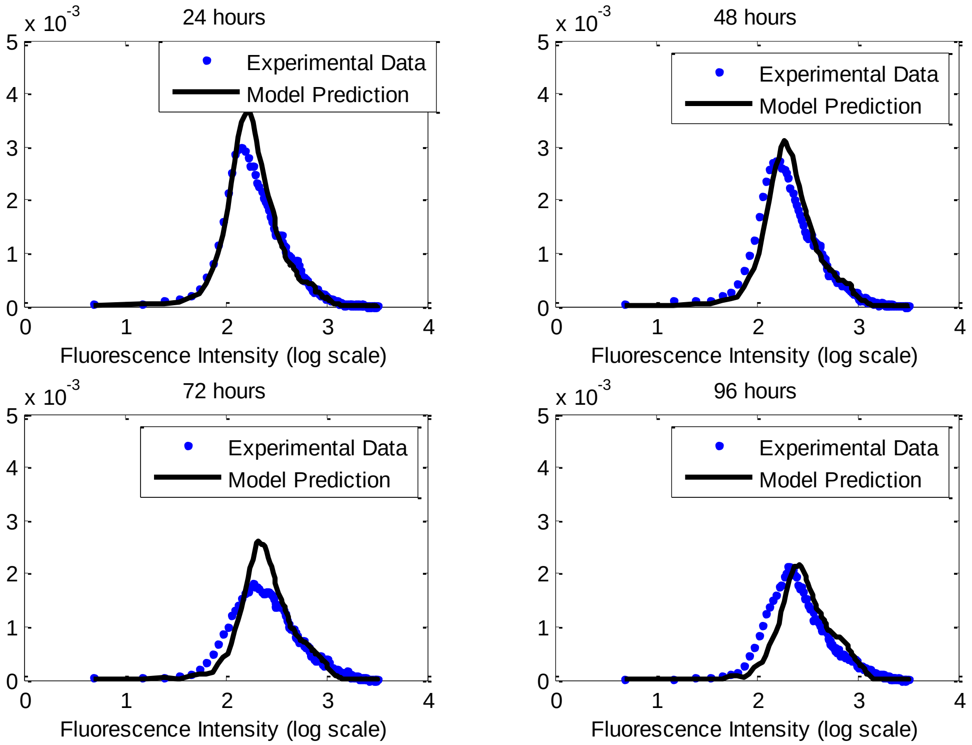


Figure 24 Plots of normalized FI distributions from experimental data and simulation of the PBM using the estimated parameters for HeLa cells stimulated with 1 μg/ml Dox.

5.4.4 Limitations in the Developed Approach

There are certain limitations in the presented approach for modeling cell populations labeled with a fluorescent reporter system. For instance, the total number of cells for each sample is not counted using flow cytometry. Thus it is challenging to estimate some of the parameters in the developed population balance model, for instance the death rate of the cells. If the total number of cells is known at each time instant where the measurement is sampled, cell number distributions instead of the normalized distributions can be used for estimating the unknown parameters. This may lead to separation of the effects of the different parameters which were previously correlated w.r.t. the normalized FI distributions, and more parameters can be estimated using experimental data.

Furthermore, for the experimental data from the HeLa Tet-on system, the parameters describing the partition function (q) and the intensity of the division rate function (a) did not have any observable effects on the resulting fluorescence intensity distributions and they may even be obsolete for the available data set (section 5.4.3). Unlike model simulations done using the nominal value of the parameters, shown in section 5.3.2, where multimodal distributions are obtained, the data obtained for the HeLa tet-on system exhibit unimodal distributions. Thus, it is possible that there is not enough "resolution" in the experimental data of this system to estimate the effects of some of the factors like the division rate of cells or the partition function parameter. Thus, because of the particular dynamics of this system, the effects of the various terms considered in the PBM (equation (5.1)) may be confounded in the resulting distributions and hard to estimate from the available experimental data.

Lastly, in previous studies (Banks et al., 2010; Luzyanina et al., 2007) involving modeling of fluorescence labeled cell populations, a fluorescent dye was used for labeling the cells which was not continuously produced inside the cells. In this work, the cell populations are labeled with a fluorescent protein which is produced inside the cells via the internal mechanisms involving gene expression and translation. These processes are inherently affected by a lot of variability (Kærn et al., 2005; Paulsson, 2005; Swain et al., 2002) and it can be challenging to incorporate the knowledge about the noise in gene expression and variability in production of fluorescent proteins in the population balance modeling framework.

5.5 Summary

In this section, a dynamic model of a cell population labeled with a fluorescent reporter system has been developed using cell population balance equation modeling technique. This model takes into account the various factors that can affect the distribution of the amount of fluorescent protein or fluorescence intensity in cells, such as the growth rate of fluorescence intensity, the division rate of cells and their partitioning behavior. This resulted in an IPDE with time and fluorescence intensity as the independent variables to model the cell number distributions. The model is simulated using an implicit finite difference scheme and the unknown parameters describing the single cell physiological parameters are estimated using experimental data.

To identify the significant parameters in the cell population balance model, a global sensitivity analysis technique called Morris method has been used. Then, the unknown parameters in the model are estimated using experimental data obtained for HeLa cells containing the Tet-on expression system using the flow cytometry technique. Two different sets of data were obtained in which the “HeLa Tet-on” cell line was stimulated with 1 μg/ml and 10 μg/ml doxycycline (Dox; the inducer for the Tet-on expression system) and the cells were monitored for a total of 4 days.

From the estimated parameters, there is an indication that higher concentrations of doxycycline cause detrimental effects on HeLa cells leading to a net lower growth rate of fluorescence intensity i.e. the rate of gene expression and a lower division rate of cells in samples stimulated with 10 μg/ml of Dox as compared to samples stimulated with 1 μg/ml of Dox.

These preliminary results demonstrate how the population balance modeling technique can be used for estimation of important physiological parameters for cell populations labeled with fluorescence reporter systems. However, there are certain limitations in the developed approach due to correlations between the effects of the unknown parameters on the normalized output fluorescence intensity distributions which makes it challenging to accurately estimate their values. Furthermore, more efficient experimental techniques as well as different reporter systems need to be investigated for complete validation of the proposed model.

6. CONCLUSIONS AND FUTURE WORK

Fluorescent reporter systems are widely used by researchers for multiple applications such as monitoring gene expression, protein-protein interactions or dynamics of signaling pathways. However, there are certain challenges that pose limitations for modeling the response obtained from these reporter systems. Some of these challenges are, for instance, these systems can consist of very large number of components with complex interactions, there can also be limited availability of experimental data both in terms of sampling points and the number of components that can be measured and the presence of large amounts of noise and variability in the response further complicates the situation. This dissertation presented several new techniques to address these challenges and illustrated them by applying them to a number of biological systems to aid in the process of mathematical modeling and estimation for systems containing fluorescent reporter systems. A brief overview of the developed techniques and the main contributions of this work are summarized below. This is followed by a discussion for future work.

6.1 Contributions

In section 3 of this dissertation, an inverse problem has been formulated to estimate the dynamics of transcription factors, which is a crucial molecule that initiates the transcription process, using experimental data of fluorescence obtained in fluorescent reporter systems. The main contributions in this study are that, unlike previous works, any complex dynamics of the transcription factor profiles can be estimated using the presented technique without any restrictions on the shapes of the estimated profiles. Furthermore, this formulation also takes into account that the experimental data could be limited and may have missing data points. Thus, regularization techniques have been incorporated in the estimation formulation to solve the underdetermined and ill-conditioned inverse problem. This formulation can be used for estimation of the dynamics of any transcription factor if the corresponding data for a fluorescent reporter is available and thus the developed procedure is not application specific. Also, this technique can generally be applied as a guideline for estimation of concentrations of unknown molecules or proteins in biological systems when it is difficult to measure them directly, however, when the measurements of a certain output is available which is directly or indirectly affected by the molecule of interest. (Bansal et al., 2012)

Section 4 presented a new experimental design criterion to facilitate the use of multiple fluorescent reporters in experiments. The major challenge in using multiple reporters together is that their emission spectra can overlap to a large extent making it difficult to separate the effect of individual reporter proteins. Thus, this procedure developed guidelines to select the fluorescent proteins to use together in an experiment such that the estimation of the contributions of the individual proteins to overall emission intensity can be determined as accurately as possible. The developed design criterion can be used for screening of proteins for applications where multiple events need to be monitored and it can be used in addition to other considerations which are currently used, such as stability, toxicity, and maturation time of fluorescent proteins. (Bansal et al., 2013)

Finally, in section 5, a population balance model has been developed to describe the dynamics of the fluorescence intensity distributions observed in fluorescent reporter systems. This model is useful in describing the various factors affecting the FI distributions such as the net growth rate of fluorescence intensity, the partitioning of the fluorescent protein on cell division and the knowledge about whether this partition is equal or unequal has also been integrated into the model. Preliminary validation results for this model are obtained by using the experimental data for HeLa cells containing the Tet-on expression system. The model has been used to compare and obtain preliminary hypothesis about the difference in the response of the HeLa Tet-on system when it is stimulated with two different concentrations of doxycycline.

6.2 Future Work

Some of the suggestions for future work and extensions of the work in this dissertation are discussed in this subsection.

The inverse problem formulation developed in Section 3 of this dissertation is evaluated by assuming that there is random Gaussian noise in the fluorescence intensity profiles. However, the noise in biological systems can be more closely characterized by simulating the ODE model describing the transcription and translation process using a stochastic simulation algorithm, for instance the ill espie’s algorithm (Gillespie, 1977). Thus, the presented inverse problem formulation should also be evaluated by using the fluorescence intensity profiles simulated using a stochastic simulation algorithm.

Secondly, in Section 4, the fluorescence intensity of a mixture of fluorescent proteins is assumed to be a linear superposition of the intensities of the reporters present in the mixture. It was also observed that the overall error in the estimated contributions of the individual reporters increased as the total number of reporters in a mixture were increased. One possible explanation of these errors is that there could be interactions effects between the intensities of different fluorescent reporters in the mixture. This can be addressed by updating the linear formulation of equation (4.2) with 2nd order interaction terms such that the overall fluorescent intensity can be represented as

$$\begin{aligned} y(x) &= \theta_1 g_1(x) + \theta_2 g_2(x) + \cdots + \theta_k g_k(x) \\ &\quad + \theta_{k+1} g_1(x) g_2(x) + \theta_{k+2} g_2(x) g_3(x) + \ldots + \theta_{k(k+1)/2} g_{k-1}(x) g_k(x) + e(x) \end{aligned} \tag{6.1}$$

Furthermore, in a generalized representation of equation (4.8), the D-optimality criterion is applied to the fisher information matrix (FIM) calculated using the sensitivity matrix (**S**) of the output of the model w.r.t. to the unknown parameters such that the formulation can be written as,

$$\begin{aligned} &\max \left[\det\left(\mathbf{FIM}\right)\right] \\ &\text{where } \mathbf{FIM} = \mathbf{S}^T \mathbf{S} \end{aligned} \tag{6.2}$$

and the sensitivity matrix **S** is calculated as

$$\mathbf{S} = \begin{bmatrix} \dfrac{\partial y(x_1)}{\partial \theta_1} & \cdots & \dfrac{\partial y(x_1)}{\partial \theta_K} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial y(x_n)}{\partial \theta_1} & \cdots & \dfrac{\partial y(x_n)}{\partial \theta_K} \end{bmatrix} \tag{6.3}$$

Thus, for the case where the overall fluorescence intensity is modeled as given in equation (6.1), the sensitivity matrix (**S)** as well as the FIM will be a function of the parameters $[\theta_1,...,\theta_k,\theta_{k+1},...,\theta_{k(k+1)/2}]$ which are not known apriori. Thus, an iterative model-based optimal design methodology (Franceschini and Macchietto, 2008) needs to be adopted for selecting the fluorescent proteins to use together. In this technique, the experimental design is first carried out using approximate or nominal values of the unknown parameters followed by conducting the experiments with that design. Then, using the data from those designed experiments the unknown parameter values are re-estimated. This process is repeated until the estimated parameters are obtained within the desired accuracy.

The next suggestion for future work involves updating the approach for validating the PBM described in section 5. The developed PBM could not be completely validated due to correlated effects of the parameters on the normalized fluorescence intensity distributions and the parameters like death rate having no effect on the normalized distributions at all. There also seems to be not enough "resolution" in the data for the HeLa Tet-on system to separate the effects of various factors affecting the FI distributions. Thus, firstly there is a need to obtain experimental data for different reporter systems and cell lines for complete validation of the proposed model. Also, obtaining additional experimental data for the total number of cells at each time instant can aid in independently estimating some of the unknown parameters in the PBM and thus reducing the limitations in model validation due to redundant effects of the parameters on normalized fluorescent intensity distributions. In addition, different functional forms for the physiological functions such as the rate of increase of fluorescence intensity, death rate or division rate should be considered such that they are relevant for the reporter system or cell line under consideration. Some of the recent works (Banks et al., 2011b; Luzyanina et al., 2007; Mantzaris, 2006) have addressed this issue in a comprehensive manner and provide cues for selection of appropriate functional forms for the cell physiological functions.

Finally, the confounding of the “multimodal features” in the fluorescence intensity distributions can be attributed to the noise in fluorescence protein production in cells due to stochastic nature of gene expression. This variability has not been considered in the population balance model developed in this work. The future work can involve combining the population balance modeling technique with stochastic descriptions of intracellular gene regulatory processes (Hasenauer et al., 2011b; Shu et al., 2012).

REFERENCES

Abu-Absi, N.R., Zamamiri, A., Kacmar, J., Balogh, S.J., Srienc, F., 2003. Automated flow cytometry for acquisition of time-dependent population data. Cytometry Part A 51A, 87-96.

Aderem, A., 2005. Systems Biology: Its practice and challenges. Cell 121, 511-513.

Aster, R.C., Borchers, B., Thurber, C., 2005. Parameter estimation and inverse problems. Elsevier Academic Press, Waltham, Massachusetts.

Banks, H., Sutton, K.L., Thompson, W.C., Bocharov, G., Roose, D., Schenkel, T., Meyerhans, A., 2011a. Estimation of cell proliferation dynamics using CFSE data. Bulletin of Mathematical Biology 73, 116-150.

Banks, H.T., Charles, F., Jauffret, M.D., Sutton, K.L., Clayton Thompson, W., 2010. Label structured cell proliferation models. Applied Mathematics Letters 23, 1412-1415.

Banks, H.T., Sutton, K.L., Thompson, W.C., Bocharov, G., Doumic, M., Schenkel, T., Argilaguet, J., Giest, S., Peligero, C., Meyerhans, A., 2011b. A new model for the estimation of cell proliferation dynamics using CFSE data. Journal of Immunological Methods 373, 143-160.

Bansal, L., Chu, Y., Laird, C., Hahn, J., 2012. Regularization of inverse problems to determine transcription factor profiles from fluorescent reporter systems. AIChE Journal 58, 3751-3762.

Bansal, L., Nelson, R., Yang, E., Jayaraman, A., Hahn, J., 2013. Experimental design of systems involving multiple fluorescent protein reporters. Chemical Engineering Science 101, 191-198.

Biosciences, B., 2000. Introduction to flow cytometry: A learning guide. Manual Part 11-11032-01, 13–14.

Borcea, L., 2002. Electrical impedance tomography. Inverse Problems 18, 99-136.

Brukilacchio, T.J., 2003. A diffuse optical tomography system combined with X-ray mammography for improved breast cancer detection. Dissertation. Tufts University, Medford, Massachusetts.

Butcher, E.C., Berg, E.L., Kunkel, E.J., 2004. Systems Biology in drug discovery. Nature Biotechnology 22, 1253-1259.

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W., Prasher, D.C., 1994. Green fluorescent protein as a marker for gene expression. Science 263, 802-805.

Chen, Y.-T., Tsai, M.-S., Yang, T.-L., Ku, A.T., Huang, K.-H., Huang, C.-Y., Chou, F.-J., Fan, H.-H., Hong, J.-B., Yen, S.-T., Wang, W.-L., Lin, C.-C., Hsu, Y.-C., Su, K.-Y., Su, I.C., Jang, C.-W., Behringer, R.R., Favaro, R., Nicolis, S.K., Chien, C.-L., Lin, S.-W., Yu, I.S., 2012. *R26R-GR*: A Cre-activable dual fluorescent protein reporter mouse. PLoS ONE 7, e46171.

Chu, Y., Hahn, J., 2012. Generalization of a parameter set selection procedure based on orthogonal projections and the D-optimality criterion. AIChE Journal 58, 2085-2096.

Cowan, S.E., Gilbert, E., Khlebnikov, A., Keasling, J., 2000. Dual labeling with green fluorescent proteins for confocal microscopy. Applied and Environmental Microbiology 66, 413-418.

Cox, R., Dunlop, M., Elowitz, M., 2010. A synthetic three-color scaffold for monitoring genetic regulation and noise. Journal of Biological Engineering 4, 10.

De Jong, H., 2002. Modeling and simulation of genetic regulatory systems: A literature review. Journal of Computational Biology 9, 67-103.

De Jong, H., Ranquet, C., Ropers, D., Pinel, C., Geiselmann, J., 2010. Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. BMC Systems Biology 4, 55.

Dickinson, M., Bearman, G., Tille, S., Lansford, R., Fraser, S., 2001. Multi-spectral imaging and linear unmixing add a whole new dimension to laser scanning fluorescence microscopy. Biotechniques 31, 1272-1279.

Dössel, O., 2000. Inverse problem of electro- and magnetocardiography: Review and recent progress. International Journal of Bioelectromagnetism 2, 2.

Evrogen, 2002. Basic fluorescent proteins. Last Accessed: 27th December, 2012. Available: http://www.evrogen.com/products/basicFPs.shtml

Fierro, R.D., Golub, G.H., Hansen, P.C., O'Leary, D.P., 1997. Regularization by truncated total least squares. SIAM Journal on Scientific Computing 18, 1223.

Fife, R., Sledge Jr, G., 1998. Effects of doxycycline on cancer cells in vitro and in vivo. Advances in Dental Research 12, 94-96.

Finkelstein, A., Hetherington, J., Li, L., Margoninski, O., Saffrey, P., Seymour, R., Warner, A., 2004. Computational Challenges of Systems Biology. Computer 37, 26-33.

Finkenstädt, B., Heron, E.A., Komorowski, M., Edwards, K., Tang, S., Harper, C.V., Davis, J.R.E., White, M.R.H., Millar, A.J., Rand, D.A., 2008. Reconstruction of transcriptional dynamics from gene reporter data using differential equations. Bioinformatics 24, 2901-2907.

Fischer, P., Lehmann, U., Sobota, R.M., Schmitz, J., Niemand, C., Linnemann, S., Haan, S., Behrmann, I., Yoshimura, A., Johnston, J.A., Muller-Newen, G., Heinrich, P.C., Schaper, F., 2004. The role of the inhibitors of Interleukin-6 signal transduction SHP2 and SOCS3 for desensitization of Interleukin-6 signalling. Biochemistry Journal 378, 449 - 460.

Franceschini, G., Macchietto, S., 2008. Model-based design of experiments for parameter precision: State of the art. Chemical Engineering Science 63, 4846-4872.

Friedrich, S., 1999. Cytometric data as the basis for rigorous models of cell population dynamics. Journal of Biotechnology 71, 233-238.

Gabriel, P., Garbett, S.P., Quaranta, V., Tyson, D.R., Webb, G.F., 2012. The contribution of age structure to cell population responses to targeted therapeutics. Journal of Theoretical Biology 311, 19-27.

Ganusov V. Bril'kov A. Pechurkin N. 2000. Mathematical modeling of population dynamics of unstable plasmid-bearing bacterial strains under continuous cultivation in a chemostat. Biophysics 45, 881-887.

Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry 81, 2340-2361.

Golub, G.H., Hansen, P.C., O'Leary, D.P., 2000. Tikhonov regularization and total least squares. SIAM Journal on Matrix Analysis and Applications 21, 185-194.

Grégori, G., Patsekin, V., Rajwa, B., Jones, J., Ragheb, K., Holdman, C., Robinson, J.P., 2012. Hyperspectral cytometry at the single-cell level using a 32-channel photodetector. Cytometry Part A 81, 35-44.

Hansen, M.C., Palmer, R.J., Udsen, C., White, D.C., Molin, S., 2001. Assessment of GFP fluorescence in cells of streptococcus gordonii under conditions of low ph and low oxygen concentration. Microbiology 147, 1383-1391.

Hansen, P.C., 1990. Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank. SIAM Journal of Scientific Computing 11, 503-518.

Hansen, P.C., 1992. Analysis of discrete ill-posed problems by means of the L-curve. SIAM Review 34, 561-580.

Hansen, P.C., 2010. Discrete inverse problems: Insight and algorithms. Society for Industrial and Applied Mathematics.

Hasenauer, J., Waldherr, S., Doszczak, M., Radde, N., Scheurich, P., Allgöwer, F., 2011a. Identification of models of heterogeneous cell populations from population snapshot data. BMC Bioinformatics 12, 1-15.

Hasenauer, J., Waldherr, S., Doszczak, M., Scheurich, P., Radde, N., Allgöwer, F., 2011b. Analysis of heterogeneous cell populations: A density-based modeling and identification framework. Journal of Process Control 21, 1417-1425.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: Data mining, inference and prediction. 2nd ed. Springer

Heath, A.P., Kavraki, L.E., 2009. Computational challenges in Systems Biology. Computer Science Review 3, 1-17.

Henson, M.A., 2003. Dynamic modeling of microbial cell populations. Current Opinion in Biotechnology 14, 460-467.

Henson, M.A., Müller, D., Reuss, M., 2002. Cell population modelling of yeast glycolytic oscillations. Biochem. J. 368, 433–446.

Hinkelmann, K., 2012. Design and analysis of experiments, Volume 3: Special designs and applications. Wiley.

Hiraoka, Y., Shimi, T., Haraguchi, T., 2002. Multispectral imaging fluorescence microscopy for living cells. Cell Structure and Function 27, 367-374.

Hjortsø, M.A., 2006. Population balances in biomedical engineering: Segregation through the distribution of cell states. McGraw-Hill Professional.

Hoffman, R.M., 2005. The multiple uses of fluorescent proteins to visualize cancer in vivo. Nature Reviews Cancer 5, 796-806.

Hu, C.D., Kerppola, T.K., 2003. Simultaneous visualization of multiple protein interactions in living cells using multicolor fluorescence complementation analysis. Nature Biotechnology 21, 539.

Huang, Z., Chu, Y., Cunha, B., Hahn, J., 2010a. Generalisation of a procedure for computing transcription factor profiles. IET Systems Biology 4, 108-118.

Huang, Z., Chu, Y., Hahn, J., 2012. Computing transcription factor distribution profiles from green fluorescent protein reporter data. Chemical Engineering Science 68, 340-354.

Huang, Z., Moya, C., Jayaraman, A., Hahn, J., 2010b. Using the tet-on system to develop a procedure for extracting transcription factor activation dynamics. Molecular Biosystems 6, 1883-1889.

Huang, Z., Senocak, F., Jayaraman, A., Hahn, J., 2008. Integrated modeling and experimental approach for determining transcription factor profiles from fluorescent reporter data. BMC Systems Biology 2, 64.

Jothi, R., Balaji, S., Wuster, A., Grochow, J.A., Gsponer, J., Przytycka, T.M., Aravind, L., Babu, M.M., 2009. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. Molecular Systems Biology 5.

Kærn, M., Elston, T.C., Blake, W.J., Collins, J.J., 2005. Stochasticity in gene expression: From theories to phenotypes. Nature Reviews Genetics 6, 451-464.

Kitano, H., 2002. Systems Biology: A brief overview. Science 295, 1662-1664.

Kretzschmar, A.K., Dinger, M.C., Henze, C., Brocke-Heidrich, K., Horn, F., 2004. Analysis of STAT3 (signal transducer and activator of transcription 3) dimerization by fluorescence resonance energy transfer in living cells. Biochemical Journal 377, 289.

Lansford, R., Bearman, G., Fraser, S.E., 2001. Resolution of multiple green fluorescent protein color variants and dyes using two-photon microscopy and imaging spectroscopy. Journal of Biomedical Optics 6, 311-318.

Leveau, J.H.J., Lindow, S.E., 2001. Predictive and interpretive simulation of green fluorescent protein expression in reporter bacteria. Journal of Bacteriology 183, 6752-6762.

Lippincott-Schwartz, J., Patterson, G.H., 2003. Development and use of fluorescent protein markers in living cells. Science 300, 87-91.

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., Gerstein, M., 2004. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431, 308-312.

Luzyanina, T., Roose, D., Schenkel, T., Sester, M., Ehl, S., Meyerhans, A., Bocharov, G., 2007. Numerical modelling of label-structured cell population growth using CFSE distribution data. Theoretical biology & Medical Modelling 4, 26.

Ma, H., Bryers, J.D., 2010. Non-invasive method to quantify local bacterial concentrations in a mixed culture biofilm. Journal of Industrial Microbiology & Biotechnology 37, 1081-1089.

Mantzaris, N.V., 2006. Stochastic and deterministic simulations of heterogeneous cell population dynamics. Journal of Theoretical Biology 241, 690-706.

Mantzaris, N.V., Liou, J.-J., Daoutidis, P., Srienc, F., 1999. Numerical solution of a mass structured cell population balance model in an environment of changing substrate concentration. Journal of Biotechnology 71, 157-174.

Mao, H., Cremer, P.S., Manson, M.D., 2003. A sensitive, versatile microfluidic assay for bacterial chemotaxis. Proceedings of the National Academy of Sciences 100, 5449-5454.

Marino, S., Hogue, I.B., Ray, C.J., Kirschner, D.E., 2008. A methodology for performing global uncertainty and sensitivity analysis in Systems Biology. Journal of Theoretical Biology 254, 178-196.

Melas, V.B., 2005. Functional approach to optimal experimental design (lecture notes in statistics). Springer-Verlag, New York, Inc.

Mhaskar, P., Hjortsø, M.A., Henson, M.A., 2002. Cell population modeling and parameter estimation for continuous cultures of saccharomyces cerevisiae. Biotechnology Progress 18, 1010-1026.

Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33, 161-174.

Mouratidis, P.X., Colston, K.W., Dalgleish, A.G., 2007. Doxycycline induces caspase-dependent apoptosis in human pancreatic cancer cells. International Journal of Cancer 120, 743-752.

Moya, C., Huang, Z., Cheng, P., Jayaraman, A., Hahn, J., 2009. Investigation of IL-6 and IL-10 signalling via mathematical modelling. IET Systems Biology 5, 15-26.

Neher, R., Neher, E., 2004. Optimizing imaging parameters for the separation of multiple labels in a fluorescence image. Journal of Microscopy 213, 46-62.

Nikos V, M., 2006. Stochastic and deterministic simulations of heterogeneous cell population dynamics. Journal of Theoretical Biology 241, 690-706.

Nowotschin, S., Eakin, G.S., Hadjantonakis, A.-K., 2009. Live-imaging fluorescent proteins in mouse embryos: Multi-dimensional, multi-spectral perspectives. Trends in Biotechnology 27, 266-276.

Onoda, T., Ono, T., Dhar, D.K., Yamanoi, A., Nagasue, N., 2006. Tetracycline analogues (doxycycline and col-3) induce caspase-dependent and-independent apoptosis in human colon cancer cells. International Journal of Cancer 118, 1309-1315.

Patterson, G., Day, R.N., Piston, D., 2001. Fluorescent protein spectra. Journal of Cell Science 114, 837-838.

Paulsson, J., 2005. Models of stochastic gene expression. Physics of Life Reviews 2, 157-175.

Perfetto, S.P., Chattopadhyay, P.K., Roederer, M., 2004. Seventeen-colour flow cytometry: Unravelling the immune system. Nature Reviews Immunology 4, 648-655.

Raser, J.M., O'Shea, E.K., 2005. Noise in gene expression: Origins, consequences, and control. Science 309, 2010-2013.

Roessel, P.v., Brand, A.H., 2002. Imaging into the future: Visualizing gene expression and protein interactions with fluorescent proteins. Nat Cell Biol 4, 15-20.

Ronen, M., Rosenberg, R., Shraiman, B., Alon, U., 2002. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. Proc Natl Acad Sci USA 99, 10555-10560.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global sensitivity analysis: The primer. Wiley-Interscience.

Shaner, N.C., Steinbach, P.A., Tsien, R.Y., 2005. A guide to choosing fluorescent proteins. Nature Methods 2, 905-909.

Shapiro, H.M., 2005. Practical flow cytometry. Wiley.com.

Sharma, G., Martin, J., 2009. Matlab®: A language for parallel computing. International Journal of Parallel Programming 37, 3-36.

Shou, G., Xia, L., Jiang, M., Wei, Q., Liu, F., Crozier, S., 2008. Truncated total least squares: A new regularization method for the solution of ecg inverse problems. IEEE Transactions on Biomedical Engineering 55, 1327-1335.

Shu, C.-C., Chatterjee, A., Hu, W.-S., Ramkrishna, D., 2012. Modeling of gene regulatory processes by population-mediated signaling: New applications of population balances. Chemical Engineering Science 70, 188-199.

Singh, A., Jayaraman, A., Hahn, J., 2006. Modeling regulatory mechanisms in il-6 signal transduction in hepatocytes. Biotechnology and Bioengineering 95, 850 - 862.

Smith, G.D., 1985. Numerical solution of partial differential equations: Finite difference methods. Oxford University Press.

Subramanian, S., Srienc, F., 1996. Quantitative analysis of transient gene expression in mammalian cells using the green fluorescent protein. Journal of Biotechnology 49, 137-151.

Swain, P.S., Elowitz, M.B., Siggia, E.D., 2002. Intrinsic and extrinsic contributions to stochasticity in gene expression. PNAS 99, 12795-12800.

Tarantola, A., 2005. Inverse problem theory and methods for model parameter estimation. Society for Industrial and Applied Mathematics.

Tsurui, N., Hattori, Hirose, Okumura, Shirai, 2000. Seven-color fluorescence imaging of tissue samples based on fourier spectroscopy and singular value decomposition. Journal of Histochemistry & Cytochemistry 48, 653-662.

Verkruysse, W., Majaron, B., Choi, B., Nelson, J.S., 2005. Combining singular value decomposition and a non-negative constraint in a hybrid method for photothermal depth profiling. Review of Scientific Instruments 76, 024301.

Villiers, G.D., McNally, B., Pike, E., 1999. Positive solutions to linear inverse problems. Inverse Problems 15, 615.

Vogel, C.R., 2002. Computational methods for inverse problems. SIAM.

Waadt, R., Schmidt, L.K., Lohse, M., Hashimoto, K., Bock, R., Kudla, J., 2008. Multicolor bimolecular fluorescence complementation reveals simultaneous formation of alternative cbl/cipk complexes in planta. The Plant Journal 56, 505-516.

Walter, E., Pronzato, L., 1990. Qualitative and quantitative experiment design for phenomenological models - a survey. Automatica 26, 195-213.

Wang, X., Errede, B., Elston, T.C., 2008. Mathematical analysis and quantification of fluorescent proteins as transcriptional reporters. Biophysical Journal 94, 2017-2026.

Watanabe, K., Saito, K., Kinjo, M., Matsuda, T., Tamura, M., Kon, S., Miyazaki, T., Uede, T., 2004. Molecular dynamics of STAT3 on IL-6 signaling pathway in living cells. Biochemical and Biophysical Research Communications 324, 1264-1273.

Wiechert, W., 2001. ^{13}C Metabolic flux analysis. Metabolic engineering 3, 195-206.

Yamada, S., Shiono, S., Joo, A., Yoshimura, A., 2003. Control mechanism of JAK/STAT signal transduction pathway. FEBS Letters 534, 190-196.

Zhdanov, M.S., 2002. Geophysical inverse theory and regularization problems. 1st ed. Elsevier Science B.V.

Zhu, G.-Y., Zamamiri, A., Henson, M.A., Hjortsø, M.A., 2000. Model predictive control of continuous yeast bioreactors using cell population balance models. Chemical Engineering Science 55, 6155-6167.

Zhu, X., Shen, J., Liu, W., Sun, X., Wang, Y., 2010. Nonnegative least-squares truncated singular value decomposition to particle size distribution inversion from dynamic light scattering data. Appl. Opt. 49, 6591-6596.

Zimmermann, T., Rietdorf, J., Pepperkok, R., 2003. Spectral imaging and its applications in live cell microscopy. FEBS letters 546, 87.

APPENDIX A

The integrals in equation (3.8) are evaluated using eigenvalue decomposition of **A.**

$$\mathbf{A} = \boldsymbol{\Sigma}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}$$

$$e^{\mathbf{A}} = \boldsymbol{\Sigma}e^{\boldsymbol{\Lambda}}\boldsymbol{\Sigma}^{-1}$$

where

$$\boldsymbol{\Lambda} = \begin{bmatrix} \mathit{\Lambda}_1 & & 0 \\ & \ddots & \\ 0 & & \mathit{\Lambda}_q \end{bmatrix}$$; $\mathit{\Lambda}_i$'s are the eigenvalues of the **A** matrix

and the columns of the **Σ** matrix are the eigenvectors.

The integrals can be evaluated as

$$\int_a^b e^{\mathbf{A}(T-\tau)} d\tau = \int_b^a e^{\mathbf{A}(T-\tau)} d(T-\tau)$$

$$= \int_{T-b}^{T-a} e^{\mathbf{A}\tau} d\tau = \boldsymbol{\Sigma}\left(\int_{T-b}^{T-a} e^{\boldsymbol{\Lambda}\tau} d\tau\right)\boldsymbol{\Sigma}^{-1}$$

$$= \boldsymbol{\Sigma}\begin{bmatrix} \int_{T-b}^{T-a} e^{\mathit{\Lambda}_1\tau} d\tau & & 0 \\ & \ddots & \\ 0 & & \int_{T-b}^{T-a} e^{\mathit{\Lambda}_q\tau} d\tau \end{bmatrix}\boldsymbol{\Sigma}^{-1}$$

$$= \boldsymbol{\Sigma}\begin{bmatrix} \frac{1}{\mathit{\Lambda}_1}\left(e^{\mathit{\Lambda}_1(T-a)} - e^{\mathit{\Lambda}_1(T-b)}\right) & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\mathit{\Lambda}_q}\left(e^{\mathit{\Lambda}_q(T-a)} - e^{\mathit{\Lambda}_q(T-b)}\right) \end{bmatrix}\boldsymbol{\Sigma}^{-1}$$

APPENDIX B

The trapezoidal rule for approximation of finite integrals is written as

$\int_a^b f(x)dx \approx \frac{f(a)+f(b)}{2}(b-a)$

The integral under consideration is $\int_z^{z_m} \Gamma(z')p(z,z')W(z',t)dz'$ and in the presented finite difference scheme the interval $[z, z_m]$ is further divided into various sub-intervals with the fixed grid spacing of Δz. Thus, its numerical approximation using trapezoidal rule can be written as,

$$\int_z^{z_m} \Gamma(z')p(z,z')W(z',t)dz' = \left(\frac{\Gamma_i^j p_{i,i} W_i^j + \Gamma_{i+1}^j p_{i,i+1} W_{i+1}^j}{2}\right)\Delta z +$$

$$\left(\frac{\Gamma_{i+1}^j p_{i,i+1} W_i^j + \Gamma_{i+2}^j p_{i,i+2} W_{i+2}^j}{2}\right)\Delta z + \cdots + \left(\frac{\Gamma_{M-1}^j p_{i,M-1} W_{M-1}^j + \Gamma_M^j p_{i,M} W_M^j}{2}\right)\Delta z$$

Since, for the beta distribution function $p_{i,i} = 0$, the above equation can be reformulated as,

$$\Delta z \sum_{i'=i+1}^{M-1} \left(\Gamma_{i'}^j p_{i,i'} W_{i'}^j + \frac{\Gamma_M^j p_{i,M} W_M^j}{2}\right)$$