A SYSTEMS BIOLOGY APPROACH TO DEVELOP MODELS OF SIGNAL TRANSDUCTION PATHWAYS

A Dissertation

by

ZUYI HUANG

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

DOCTOR OF PHILOSOPHY

August 2010

Major Subject: Chemical Engineering

A SYSTEMS BIOLOGY APPROACH TO DEVELOP MODELS OF SIGNAL TRANSDUCTION PATHWAYS

A Dissertation

by

ZUYI HUANG

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

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Juergen Hahn
Committee Members,	Aniruddha Datta
	Mahmoud El-Halwagi
	Arul Jayaraman
	Carl Laird
Head of Department,	Michael Pishko

August 2010

Major Subject: Chemical Engineering

ABSTRACT

A Systems Biology Approach to Develop Models

of Signal Transduction Pathways. (August 2010)

Zuyi Huang, B.Eng., Tsinghua University;

M.S., Tsinghua University

Chair of Advisory Committee: Dr. Juergen Hahn

Mathematical models of signal transduction pathways are characterized by a large number of proteins and uncertain parameters, yet only a limited amount of quantitative data is available. The dissertation addresses this problem using two different approaches: the first approach deals with a model simplification procedure for signaling pathways that reduces the model size but retains the physical interpretation of the remaining states, while the second approach deals with creating rich data sets by computing transcription factor profiles from fluorescent images of green-fluorescent-protein (GFP) reporter cells.

For the first approach a model simplification procedure for signaling pathway models is presented. The technique makes use of sensitivity and observability analysis to select the retained proteins for the simplified model. The presented technique is applied to an IL-6 signaling pathway model. It is found that the model size can be significantly reduced and the simplified model is able to adequately predict the dynamics of key proteins of the signaling pathway.

An approach for quantitatively determining transcription factor profiles from GFP

reporter data is developed as the second major contribution of this work. The procedure analyzes fluorescent images to determine fluorescence intensity profiles using principal component analysis and K-means clustering, and then computes the transcription factor concentration from the fluorescence intensity profiles by solving an inverse problem involving a model describing transcription, translation, and activation of green fluorescent proteins. Activation profiles of the transcription factors NF-κB, nuclear STAT3, and C/EBPβ are obtained using the presented approach. The data for NF-κB is used to develop a model for TNF-α signal transduction while the data for nuclear STAT3 and C/EBPβ is used to verify the simplified IL-6 model.

Finally, an approach is developed to compute the distribution of transcription factor profiles among a population of cells. This approach consists of an algorithm for identifying individual fluorescent cells from fluorescent images, and an algorithm to compute the distribution of transcription factor profiles from the fluorescence intensity distribution by solving an inverse problem. The technique is applied to experimental data to derive the distribution of NF-κB concentrations from fluorescent images of a NF-κB GFP reporter system.

To my parents, my wife, and my son

ACKNOWLEDGEMENTS

My deepest gratitude is extended to my advisor, Dr. Juergen Hahn. He always provided insightful and constructive suggestions on my research and gave me the freedom to explore on my own. His belief in me encouraged me to enjoy my research even when I was struggling to solve some challenging problems. His dedication and passion for research also inspired my enthusiasm for hard work. He gave me many chances to improve my skills that will be helpful for my future career. He also cares about a student's life and family. I really appreciate all his help with me and my family. He is an ideal model as an advisor. It is a great fortune for me to work with him.

I am deeply grateful to Dr. Arul Jayaraman for his help in preparing the manuscripts of those co-authored journal papers. I benefited a lot in my research from his suggestions and experience on biological experimentation. All experimental data used in this dissertation were provided by his group. The experience of working with an excellent biological scientist like him was an asset for me.

I appreciate my other committee members, Dr. Aniruddha Datta, Dr. Mahmoud El-Halwagi, and Dr. Carl Laird for all their kindness and patience in helping me to complete this work. I would also like to acknowledge Dr. Roland R. Kaunas from the Department of Biomedical Engineering for his collaboration on one journal paper, although it is not included in this dissertation.

I would like to thank the following members of Dr. Arul Jayaraman's group, Fatih Senocak, Colby Moya, and Peng Cheng for their hard work on carrying out the

experiments related to this work. I am also thankful to my friends and fellow graduate students from the Process System Engineering group, to name a few but not limited to, Yunfei Chu, Cheryl Qu, Yu Zhu, Mitch Serpas, Abdullah Bin Mahfouz, and Eman Tora, for making me have a good time at Texas A&M University. I want to especially thank Yunfei Chu for his collaboration in writing a couple of journal papers. I am also grateful to my American friend Kelly Davidson for helping me and my family adjust to living here.

Most importantly, I would like to express my heart-felt gratitude to my parents, my wife, Qian Jia, for their love, support and patience all these years. A special thank you is given to my seven-month-old son, Cody Kangqi Huang, for being good during my writing of this dissertation. To my family I dedicate my dissertation.

NOMENCLATURE

APR	Acute Phase Response
EMP	Embden-Meyerhof-Parnas
FIM	Fisher Information Matrix
GFP	Green Fluorescent Protein
IL-6/10	Interleukin 6/10
LQR	Linear Quadratic Regulator
NAFLD	Non-alcoholic Fatty Liver Disease
ODE	Ordinary Differential Equation
PCA	Principal Component Analysis
PDE	Partial Differential Equation
RE	Response Element
SNR	Signal Noise Ratio
SBML	Systems Biology Makeup Language
TF	Transcription Factor
TNF-α	Tumor Necrosis Factor α

TABLE OF CONTENTS

Page

Page

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

1. INTRODUCTION

1.1 Signal Transduction Pathway

In biological systems, cellular networks can be categorized into the following three groups: signal transduction pathways (Bandhyopadhyay et al., 2007; Eungdamrong and Iyengar, 2004; Hoffmann et al., 2007; Pokholok et al., 2006), metabolic networks (Guimerà and Amaral, 2005; Jeong et al., 2000; Schuster et al., 2000), and gene regulatory networks (Hasty et al., 2001; Karlebach and Shamir, 2008; Paulsson, 2004). This classification is based on the different hierarchical and organizational levels of networks. Specifically, signal transduction pathways mainly deal with how the extracellular stimulus is conveyed from the membrane of the cell into the cell nucleus for target gene activation via activation or deactivation of signaling proteins in ordered sequences of biochemical reactions. Metabolic networks specify the conversions between small biochemical molecules (the metabolites) to provide the biomass and energy that are critical for the cell growth, whereas gene regulatory networks describe how some specific genes are expressed by investigating the interaction between proteins and DNA (Assmus et al., 2009). For the purpose of illustration, Fig. 1 shows one example for each of these three biological networks. Although these three networks have different components and biological functions, they work with each other to maintain the function and metabolism of the cell. For example, metabolic networks provide energy

---

This dissertation follows the style of Biotechnology and Bioengineering.

for biochemical reactions in signal transduction pathways, and the expressed proteins from signaling pathways may act as transcription factors in gene regulatory networks. While a certain amount of experimental data for metabolic and gene regulatory networks is available, only a limited amount of quantitative data about signaling proteins in signal transduction pathways can generally be obtained. Furthermore, information about transient dynamics is required for signal transduction pathways whereas steady state analysis is extensively implemented for the other two network types. As a result of these factors, mathematical modeling for signal transduction pathways is quite challenging. This dissertation mainly focuses on the modeling issues of signal transduction pathways.

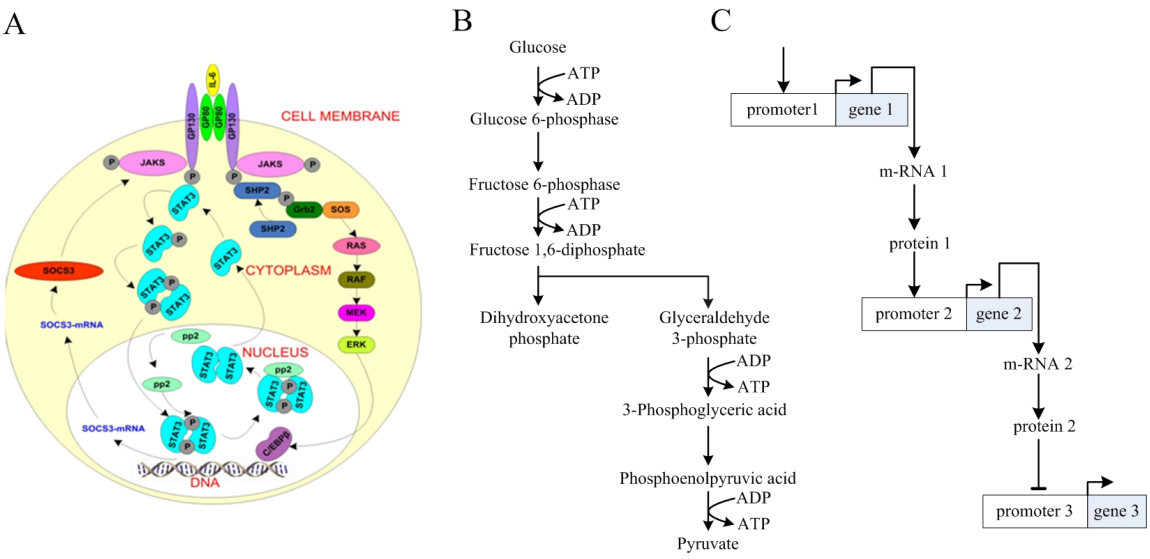


Figure 1. Biological networks. (A) IL-6 signaling pathway as an example of signal transduction pathways. SOCS3, among other proteins, is expressed due to the extracellular stimulation of IL-6 (Huang et al., 2009a). (B) the Embden-Meyerhof-Parnas (EMP) pathway as an example of metabolic networks. Nutrition resource glucose is converted to pyruvate and the released energy is shuttled in ATP (Bailey and Ollis, 1986). (C) an example of gene regulatory networks. Protein 1 activates gene 2 by binding to its promoter. This results in the expression of protein 2, which in turn inhibits gene 3 by blocking the binding sites in promoter 3.

Signal transduction pathways generally consist of cascades of biochemical reactions linking the extracellular stimulation to the target gene in the nucleus. Although different signaling pathways have different regulatory mechanisms, they share a similar framework. In the following text, the IL-6 signal transduction pathway (Fig. 1A) is used to illustrate the main features of signal transduction pathways:

1) Since IL-6 is one of the systemic inflammatory mediators involved in the regulation of the hepatic acute phase response (APR), IL-6 is secreted by the resident macrophages around the wound.
2) Following the blood stream, IL-6 arrives at the liver and stimulates liver cells by binding to its receptor at the cell membrane, resulting in the formation of a receptor complex.
3) The receptor complex auto-phosphates, activates STAT3C in the Jak-STAT pathway, and also initiates Erk-C/EBPβ pathway by binding to SHP2.
4) The activated proteins from both Jak-STAT and Erk-C/EBPβ pathways in turn activate other proteins. The signal is transduced via cascades of reactions in the cytoplasm. The activated transcription factor complex $STAT3C^{*}$-$STAT3C^{*}$ is formed in the cytoplasm.
5) The transcription factor $STAT3C^{*}$-$STAT3C^{*}$ translocates into the nucleus.
6) Transcription factor $STAT3N^{*}$-$STAT3N^{*}$ induces transcription and translation of some proteins. For example, protein SOCS3 is expressed.
7) The expressed proteins may change the function of the cell or play an important role to maintain the function of the cell. In this example, SOCS3

inhibits both Jak-STAT and Erk-C/EBPβ pathways by competing for the same binding sites with STAT3C and SHP2 in the receptor complex.

8) Some components such as nuclear STAT3 and C/EBPβ are involved in the regulation of APR.

It can be seen from the above discussion that signal transduction pathways can regulate many cellular processes (e.g., gene expression for the target proteins) and are also involved in extracellular communication. An understanding of signal transduction mechanisms offers the potential for improved treatment options for diseases. As mentioned in the above discussion, an improved understanding of IL-6 signaling pathway is helpful for improving the treatment of APR. Non-alcoholic fatty liver disease (NAFLD) is used as another example to address this point. NAFLD is a metabolic liver disease that encompasses a range of conditions from hepatic steatosis (or simple accumulation of fat in hepatocytes) to end-stage liver diseases such as fibrosis/cirrhosis (Clark 2006; McClain et al., 2004; Méndez-Sánchez et al., 2007; Sanyal 2005). It is the most common liver disease worldwide and has been estimated to affect more than 30 million people in the U.S.A. and is strongly associated with different aspects of the Metabolic Syndrome. The accumulation of fat in hepatocytes (or steatosis) is the initial step that is common to all aspects of the disease, including steatohepatitis (Lalor et al., 2007). Even though steatosis is defined as a clinically benign condition as it is reversible and does not lead to liver damage by itself, its progression to steatohepatitis results in strongly adverse effects to liver health. Therefore, it is important to improve the understanding of regulatory mechanisms involved in steatosis. It is becoming evident

that pro-inflammatory cytokines such as TNF-α or IL-6 contribute to disease progression (Diehl, 2004; Lalor et al., 2007; Rahman et al., 2007). Indeed, circulating levels of IL-6 and its soluble receptor are increased in subjects with NAFLD as compared to normal subjects (Abiru et al., 2006). Furthermore, IL-6 is also the primary mediator of hepatic insulin resistance (Videla et al., 2006; Willner et al., 2001), which further underscores the importance of IL-6 in steatosis. As shown in Fig. 1A, IL-6 signaling in the liver can occur through two pathways - the prototypical Jak-STAT pathway and the Erk-C/EBPβ pathway - to activate the transcription factors STAT3 and C/EBPβ, respectively (Heinrich et al., 1998). The Jak-STAT and Erk-C/EBPβ pathways do not function independently as they essentially compete for binding to the same IL-6 receptor complex (Heinrich et al., 1998), and STAT3 activation can be inhibited by signaling through the Erk pathway (Sengupta et al., 1998). Interestingly, the Jak-STAT and Erk-C/EBPβ pathways are also utilized differently during steatotis. STAT3 activation has been shown to decrease hepatic steatosis in obese mice both by increasing fatty acid oxidation (Hong et al., 2004) (through PPARα up-regulation) and by decreasing fatty acid synthesis (through inhibition of SREBP1c) (Inoue et al., 2004; Ueki et al., 2004). On the other hand, recent knockout mice studies have demonstrated that activation of the transcription factor C/EBPβ through the Erk1/2 (MAP kinase) pathway increases steatosis by promoting fatty acid synthesis and lipid accumulation (Rahman et al., 2007). Together, these studies suggest that the extent of steatosis can be significantly influenced by whether IL-6 signaling occurs through the Jak-STAT and/or Erk-C/EBPβ pathways. Therefore, understanding IL-6 signaling dynamics is important for developing

approaches to counter the effects of aberrant IL-6 signaling in steatotic complications.

1.2 Signal Transduction Pathway Modeling

The investigation of signal transduction pathways is one of the central themes in Systems Biology as signal transduction pathways play an important role for biological systems. However, analyzing signal transduction pathways is far from trivial as the time constants of the dynamics exhibited by proteins in the pathways can vary significantly from one protein to another, multiple pathways can be involved in signal transduction initiated by one stimulus, and crosstalk exists between signal transduction pathways both for signal transduction by the same stimulus but also for cases where the transduction was initiated by different stimuli. Furthermore, it is becoming evident (Kholodenko, 2006) that the dynamic behavior of some proteins, such as transcription factors, have a direct effect on the response of a cell to a stimulus and that only analyzing the steady state behavior is insufficient for characterizing the response. A conclusion derived from these observations is that a detailed characterization involving models of signal transduction activity is required for fully understanding the effect that stimuli, and how they interact, have on the cellular response. One possibility for developing an understanding of the dynamics of signal transduction pathways is the derivation of models describing the pathways.

Mathematical modeling for signal transduction pathways has attracted strong interest in the community of Systems Biology. Among those popular approaches extensively used are, to name some but not limited to, Boolean networks, Fuzzy inference, Bayesian Networks, and ordinary differential equations (Ideker and Lauffenburger, 2003).

1.2.1 Boolean Networks

Boolean networks consist of binary state nodes (Kauffman 1969) and are used extensively to construct gene regulatory networks from gene expression data (Karlebach and Shamir, 2008; Shmulevich et al., 2002). Recently, Boolean networks have been applied to predict the profiles of components in signal transduction pathways (Klamt et al., 2006; Klamt et al., 2007; Saez-Rodriguez at al., 2007; Wittmann et al., 2009). The procedure to construct Boolean networks for signal transduction pathways includes the following steps: first, a truth table is built to describe the state transition for all possible state values; second, Boolean functions are extracted from the truth table. The relationship between components of the pathway is then expressed as a linguistic rule that is linked by elementary ‘‘AND’’, ‘‘OR’’, and ‘‘NOT’’ gates. Finally, the state values at each point in time are inferred from the values at the previous time point based on Boolean functions. A simple signal transduction pathway is shown in Fig. 2A to illustrate the steps of developing Boolean models. Since each of the input and state variables has two values (e.g., “ON” and “OFF”), there are eight possible scenarios in the truth table (Fig. 2B). Two Boolean functions are derived from the truth table (Fig. 2C). Based on the truth table or Booleans functions, the values of states can be inferred step by step over time from Eq.(1.1)

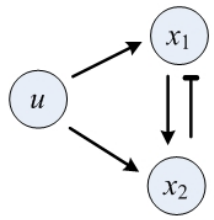
$$x(k+1) = f(x(k), u(k)) \tag{1.1}$$

where the state value of x at time $t(k+1)$ is inferred from the information at time $t(k)$.

The advantage of Boolean networks is that prior knowledge or qualitative information can be incorporated into the Boolean function and the interactions among

the components of the network (Li et al., 2006; Thakar et al., 2007). However, two drawbacks for applying Boolean networks to model signal transduction pathways might counteract the advantage: first, it is unrealistic to represent all biological signals with "ON" and "OFF" states; secondly, possible scenarios in the truth table expand exponentially with the number of components in the signaling pathway, and signal transduction pathways generally are comprised of hundreds of signaling proteins.

A



B

Input and States at $t(k)$			States at $t(k+1)$	
u	x_1	x_2	x_1	x_2
0	0	0	0	0
0	0	1	0	0
0	1	0	0	0
0	1	1	0	0
1	0	0	1	1
1	0	1	0	1
1	1	0	1	1
1	1	1	0	1

C

$x_1 = \text{NOT } x_2 \text{ AND } u$
$x_2 = u$

Figure 2. An example to illustrate the procedure of developing Boolean models. (A) A simple example of Boolean networks, (B) the truth table of all possible scenarios, (C) Boolean functions derived from the truth table.

1.2.2 Fuzzy Inference

Approaches using Fuzzy models (Mamdani and Assilian, 1975; Sugeno, 1985; Zadeh, 1965) can be regarded as an extension of traditional Boolean approaches (Davidov et al., 2003), as the probability of a state being equal to a value other than only "0" and "1" is described by the membership function. Many systems investigated in Systems Biology are characterized by a lack of quantitative data, yet a large amount of semi-quantitative data about protein concentrations in signal transduction pathways is generally available in the form of Western blots. Fuzzy models can make use of this situation as fuzzy rules

can be based upon the qualitative information that is found in the literature whereas training of the model can be performed with data that is available. The use of fuzzy logic models has been explored for modeling signal transduction pathways (Aldridge et al., 2009; Huang et al., 2009b). In the following text, the Fuzzy model presented in Huang et al., 2009b, is used as an example to describe the Fuzzy inference system for signaling pathways.

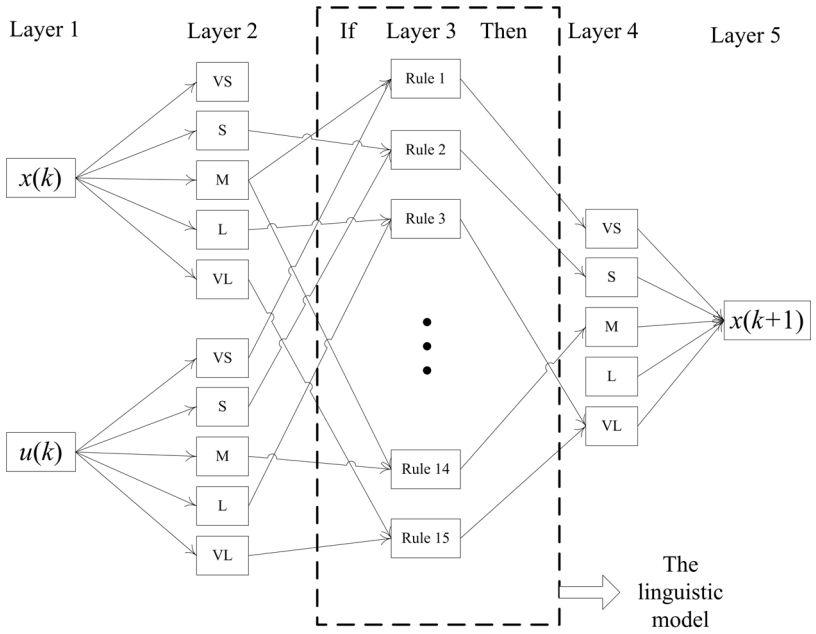


Figure 3. Fuzzy layer structure. Layer 1 shows the input and state variable at time k; Layer 2 consists of the premise membership functions that appear in the IF-part of the rule; Layer 3 is comprised of the "IF-THEN" linguistic rules; Layer 4 contains the consequent membership functions appearing in the THEN-part; Layer 5 determines the state variable at time k+1. Layer 1 and 2 are used for fuzzification while Layer 4 and 5 are applied for de-fuzzification.

Following Eq. (1.1), Fuzzy models describing dynamic processes compute the states $x(k+1)$, at a time $k+1$, from the information of the states $x(k)$ and inputs $u(k)$, at time k. In this case, $f(\cdot)$ in Eq. (1.1) is a fuzzy model with the structure shown in Fig. 3. The

values of the inputs, $x(k)$ and $u(k)$, and of the outputs, $x(k+1)$, can be assigned linguistic labels, e.g., 'Very Small' (VS), 'Small' (S), 'Medium' (M), 'Large' (L), and 'Very Large' (VL). Linguistic rules can be formulated that connect the linguistic labels for $x(k)$ and $u(k)$ via an "IF" condition with a "THEN" part that determines the resulting linguistic label for $x(k+1)$. Specifically, the fuzzy rules are of the form:

IF ($x(k)$ is $A_{x(k)}$) AND ($u(k)$ is $A_{u(k)}$), THEN ($x(k+1)$ is $A_{x(k+1)}$),

where $A_{x(k)}$, $A_{u(k)}$ and $A_{x(k+1)}$ are the linguistic labels for $x(k)$, $u(k)$, and $x(k+1)$, respectively, generated for the data points.

One example form of these membership functions in Layer 2 and 4 is shown in Eq.(1.2)

$$\mu_A(x) = \exp(-\frac{(x-c)^2}{\sigma^2}) \tag{1.2}$$

where μ_A refers to the degree to which x belongs to the linguistic label A, c represents the center of the membership functions and σ determines the width of the membership functions. The output of each node in layer 2 is the output from the corresponding membership function as given by equation (1.2). The output of the node in layer 3 is the smallest value of the inputs to that node. The output of layer 4 is the largest value of the inputs to that node. The output of layer 5 is calculated by:

$$O_i^{(5)} = \frac{\sum_{j=1}^{5} O_{i,j}^{(4)} \sigma_{i,j}^{(4)} c_{i,j}^{(4)}}{\sum_{j=1}^{5} O_{i,j}^{(4)} \sigma_{i,j}^{(4)}} \tag{1.3}$$

where $O_{i,j}^{(4)}$ refers to the output of the node in layer 4, which connects to node i in layer 5

and represents linguistic label j, where j, equal to 1,2,…,5, represents the five linguistic labels; $\sigma_{i,j}^{(4)}, c_{i,j}^{(4)}$ represent the parameters of the membership function of node $O_{i,j}^{(4)}$.

1.2.3 Bayesian Networks

Bayesian networks have been extensively used for modeling gene regulatory networks from steady state data (Friedman et al., 2000; Hartemink et al., 2001; Yu et al., 2004). For example, the probability of gene a_2 being expressed under the condition that gene a_1 is expressed can be calculated by the conditional probability:

$$\Pr(a_2 \mid a_1) = \frac{\Pr(a_1 \mid a_2) \cdot \Pr(a_2)}{\Pr(a_1)} \tag{1.4}$$

where $\Pr(a_2|a_1)$ is the posterior density function, $\Pr(a_1|a_2)$ refers to the likelihood function, and $\Pr(a_2)$ is the prior density function. Fig. 4 gives a simple example for gene regulatory networks that is used to illustrate the inference of Bayesian networks.

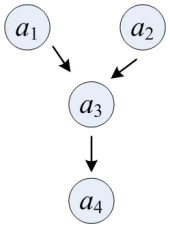


Figure 4. An example for gene regulatory networks. Gene a_3 depends on the co-expression of gene a_1 and a_2 while gene a_4 only depends on the expression of gene a_3 (Klipp et al., 2005).

The probability of all genes being expressed is calculated as:

$$\Pr(\boldsymbol{\alpha}) = \prod_i \Pr(a_i \mid L(a_i)) \tag{1.5}$$

where $\boldsymbol{\alpha}$ is the vector containing the expressed status for all genes, and $L(a_i)$ contains the parent variables that directly regulate the expression of gene a_i. For the example shown in Fig. 4, the probability of all genes being expressed is equal to:

$$\Pr(a_1, a_2, a_3, a_4) = \Pr(a_1) \cdot \Pr(a_2) \cdot \Pr(a_3 \mid a_1, a_2) \cdot \Pr(a_4 \mid a_3) \tag{1.6}$$

Bayesian inference obeys the Markov assumption, that is, each gene only conditionally depends on its parents and is independent of its other ancestors. Therefore, Bayesian approach is not applicable for the situation where mutual dependencies (e.g. cycles) between the components of networks exist.

Bayesian inference can be applied for modeling the dynamics of signal transduction pathways by reforming the conditional probability in Eq. (1.4) to the form given by Eq. (1.7). Bayesian networks have been used for modeling signal transduction pathways (Purutcuoglu and Wity, 2008; Sachs et al. 2002; Sachs et al., 2005). In Bayesian networks, the qualitative information can be incorporated into the structure of the network, the likelihood function and the prior probability distribution of the Bayes' rules (Chang et al., 2008; Lucas, 2005). One drawback of Bayesian networks is that the prior information may not always be available, though.

$$\Pr(x(t+1) \mid x(t)) = \frac{\Pr(x(t) \mid x(t+1)) \cdot \Pr(x(t+1))}{\Pr(x(t))} \tag{1.7}$$

1.2.4 Ordinary Differential Equations (ODEs)

While Boolean networks and Bayesian networks are popular for deriving gene regulatory networks from steady state data, ODE models are the most common approach for describing dynamic properties of signal transduction pathways. In this approach, the

profiles of the components in signaling pathways can be quantitatively predicted from the model. Furthermore, the interaction between the components can be quantitatively investigated. Equation (1.8) is a general form of an ODE model. A more commonly used form of the ODE model for signaling pathways is represented in Eq. (1.9).

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, \mathbf{p}, \mathbf{u}) \tag{1.8}$$

where $\mathbf{x} \in R^{n_x}$ is a vector containing concentrations/amounts of the states of the model, $\mathbf{p} \in R^{n_p}$ is a vector of the parameters, and $\mathbf{u} \in R^{n_u}$ is the vector of inputs of the system.

$$\frac{dx_i}{dt} = \sum v_{i,produced} - \sum v_{i,consumed} \tag{1.9}$$

where the left term of the equation means the rate of change of component x_i, the two terms of the right side respectively represent the rate of x_i formed and consumed in all reactions. Equation (1.9) is based on the law of mass conservation, that is, the rate of change of a component depends on the production rate of the component as well as the consumption rate. In this case, the cell is usually considered as a batch reactor where exchange between the cell and its surrounding environment is limited. The rates $v_{i,produced}$ and $v_{i,consumed}$ are derived from the chemical kinetics of the associated reactions. Most chemical reactions are described by the mass action kinetics while some other chemical reactions are represented by Michaelis-Menten kinetics. One example for Michaelis-Menten kinetics is the process of a gene in DNA transcribed to mRNA due to the binding of transcription factors to the promoter of the gene. The enzyme reaction of invertase (Brown, 1902) is used as an example for the derivation of ODE models based upon the mass action kinetics and Michaelis-Menten kinetics.

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + G \tag{1.10}$$

where E is the free enzyme, S is the substrate, ES the enzyme-substrate complex, and G is the product. S is consumed in the forward reaction where E and S react to form ES. Based on the mass action kinetics, the consumption rate of S is proportional to the concentrations of reactants S and E as well as the forward rate constant k_1. Similarly, the production rate of S depends on the concentration of ES and the backward rate constant k_{-1}. Therefore, the rate of change of S is determined as:

$$\frac{dC_s}{dt} = v_{S,produced} - v_{S,consumed} = k_{-1}C_{ES} - k_1 C_E C_s \tag{1.11}$$

where C_s, C_E, and C_{ES} are the concentrations of S, E, and ES respectively. The ODEs for other components can be derived in the same way.

$$\frac{dC_{ES}}{dt} = k_1 C_E C_s - (k_{-1} + k_2)C_{ES} \tag{1.12}$$

$$\frac{dC_E}{dt} = (k_{-1} + k_2)C_{ES} - k_1 C_E C_s \tag{1.13}$$

$$\frac{dC_G}{dt} = k_2 C_{ES} \tag{1.14}$$

where C_G is the concentration of product G. Based upon the assumptions that the formation of ES from E and S and vice versa is much faster than the decomposition process of ES into E and G (i.e., k_1, $k_{-1} >> k_2$), and that ES reaches a quasi-steady state, Eq. (1.10) can be simplified as:

$$S \longrightarrow G \tag{1.15}$$

where the rate v is expressed as Michaelis-Menten kinetics

$$v = \frac{V_{\max} \cdot C_s}{K_m + C_s} \tag{1.16}$$

where $V_{\max}$ is the maximal velocity and K_m is Michaelis constant.

1.2.5 Review of Some Popular Models of Signal Transduction Pathways

Modeling of signal transduction pathways has attracted significant interest in the Systems Biology community. Some databases for signal transduction pathways are available, such as KEGG (www.genome.jp/kegg/pathway.html), BioCarta (www.biocarta.com), Reactome (www.reactome.org), BioCyc (www.biocyc.org), MetaCyc (www.metacyc.org), PID (Pathway Interaction Database: pid.nci.nih.gov), and Pathguide (www.pathguide.org) (Assmus et al., 2009). Mathematical models written in the Systems Biology Makeup Language (SBML) can be downloaded from the website (http://www.ebi.ac.uk/biomodels-main/). Among those extensively investigated signal transduction pathways is EGFR signaling (Oda et al., 2005). Specifically, several ODE models have been developed for the sub-pathways of EGFR signaling, e.g., MAPK pathways (Brightman et al., 2000; Kholodenko et al., 1999; Orton et al., 2005; Schoeberl et al., 2002). As discussed in subsection 1.1, an improved understanding of regulatory mechanisms underlying IL-6 signaling and TNF-α signaling could provide potential treatment options for the diseases like steatosis. Due to this the focus of this dissertation is on modeling these two signal transduction pathways.

A significant amount of information has been presented in the literature on IL-6 signal transduction including the structure of the signal transduction pathway and qualitative information in the form of Western blots (Fasshauer et al., 2004; Heinrich et

al., 2003; Lang et al., 2003). However, only a limited number of fundamental models (Huang et al., 2007; Singh et al., 2006; Yamada et al., 2003) exist due to the limited amount of quantitative data which leads to these models containing a large number of uncertain parameters. Similarly, the main components of the TNF-α stimulated NF-κB signaling pathway have also been extensively studied and several mathematical models have been presented (Hoffman et al., 2002; Lipniacki et al., 2004; Rangamani and Sirovich, 2007). Although a certain amount of qualitative data in the form of Western blots exits for proteins involved in this pathway (Hoffman et al., 2002; Lee et al., 2000), very limited quantitative data are available.

1.3 Motivation of This Research

Deriving an accurate signal transduction pathway model is non-trivial as the mechanisms tend to involve many components and the system will have a large degree of uncertainty in both its structure and parameter values. Validation and refinement of any model is a crucial step for modeling signal transduction pathways. However, validation of the available models is challenging because of two points: (1) the available experimental data are limited; (2) the identifiability of these models is generally low due to the large number of uncertain parameters in the model and interaction among the components of the signal transduction network. Therefore, the availability of quantitative data plays an important role for reducing uncertainties of signaling structure and parameter values, however, it is non-trivial to derive a significant amount of appropriate data. Transcription factor data are especially important as it is becoming evident that the dynamic behavior of transcription factors has a direct effect on the

response of cells to stimuli (Kholodenko, 2006). One way to address point (1) is to develop techniques for deriving quantitative data of transcription factors.

One option to address point (2) is to perform sensitivity analysis of parameters. Only parameters that are found to be important need to be estimated from data and all other parameters are set to their nominal values. While this approach can result in models that provide a good fit for experimental data, it has the drawback that one has to deal with what is essentially an overparameterized model. Model reduction/simplification is an alternative for addressing this problem, as a smaller number of parameters/states may be appropriate if only a limited amount of experimental data is available.

This dissertation addresses the above-mentioned two challenging issues for model verification using two different approaches: techniques for deriving quantitative data of transcription factors are investigated, and a model simplification approach that retains the physical interpretation of the remaining states and parameters is developed. A detailed literature survey on these two approaches is performed in the next subsection.

1.4 Literature Survey

Systems Biology seeks to develop models for describing cellular behavior on the basis of regulatory molecules such as transcription factors and signaling kinases. The control of gene expression by transcription factors is an integral component of cell signaling and gene expression regulation (Corvinus et al., 2005; Judd et al., 2004). Different transcription factors exhibit different expression and activation dynamics, and together govern the expression of specific genes and cellular phenotypes (Heinrich et al., 2003). An important requirement for the development of these signal transduction

models is the ability to quantitatively describe the activation dynamics of transcriptions so that parameters can be estimated for model development. One popular approach for collecting experimental data of transcription factors involves Western blotting (e.g. in Birtwistle et al., 2007 and Hoffmann et al., 2002). While performing a Western blot is a relatively simple experiment, it does have the drawbacks that (a) Western blotting is a destructive measurement technique, and (b) the data is semi-quantitative in nature (Kurien and Scofield, 2006; Pan et al., 2006). The first drawback poses a problem for the use of Western blots for experiments where a time series of a concentration profile of a particular protein is to be measured while the latter results from the limitation of the technique itself, i.e., it is not always possible to determine "how black a Western blot is" and what protein concentration this level of color corresponds to. The activation of transcription factors under different conditions can be also monitored using electrophoretic mobility shift assay or chromatin immunoprecipitation (Elnitski et al., 2006). Similar to the Western blotting approach, these techniques provide snapshots of activation at a small set of single time points, yielding only qualitative or semi-quantitative data at best. Dynamics of transcription factors are not captured in these approaches due to limited sampling points and frequencies. Hence, these methods are not ideal for investigating time-dependent activation of transcription factors in a quantitative manner.

More recently, fluorescence-based reporter systems have been developed for the continuous and non-invasive monitoring of transcription factors and the elucidation of regulatory molecule dynamics. Recent studies (King et al., 2007; King et al., 2008;

Thompson et al., 2004; Wieder et al., 2005) have used green fluorescent protein (GFP) as a reporter molecule for continuously monitoring activation of a panel of transcription factors, underlying the inflammatory response in hepatocytes for 24 hours. These systems involve expressing GFP under the control of a minimal promoter such that GFP expression and fluorescence is observed only when a transcription factor is activated (i.e., when the transcription factor binds to its specific DNA binding sequence and induces expression from a minimal promoter) (Figure 5A and 5B). The dynamics of GFP fluorescence is used as the indicator for dynamics of the transcription factor being profiled. The primary drawback with this approach is that it does not provide direct activation rates of the transcription factors being investigated. Even though transcription factor dynamics influence GFP dynamics, the relationship between the two is non-trivial as the induction of GFP fluorescence itself involves multiple steps (i.e., transcription of GFP mRNA, GFP protein translation, post-translational processing, etc) (Subramanian and Srienc, 1996), and not all of these steps contribute equally to regulation of GFP expression. The observed fluorescence dynamics in GFP reporter cell systems is the result of two different dynamics: (1) the dynamics of transcription factor activation by a soluble stimulus-mediated signal transduction pathway and (2) the dynamics of GFP expression, folding, and maturation. Therefore, it is necessary to uncouple the effects of these independent systems in order to quantitatively determine transcription factor activation profiles underlying cellular phenotypes. Developing a model to describe GFP dynamics is a potential solution.

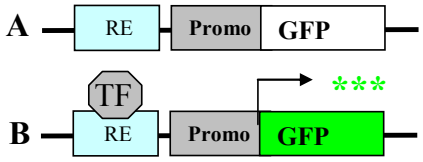


Figure 5. GFP-based reporter systems for investigating transcription factor (TF) activation. The DNA response element (RE) to which the TF binds is upstream of a minimal promoter that controls GFP expression. (A) No fluorescence is observed in the absence of TF binding, (B) Binding of TF leads to promoter activation and GFP fluorescence.

The observed fluorescence dynamics has to be quantified first for deriving transcription factor activation profiles. An automated image analysis procedure to identify the GFP localization regions with standard MATLAB commands has been presented in reference (Venkataraman et al., 2005), however, the procedure only determines regions of fluorescence and does not provide quantitative data about the fluorescence intensity. Analyzing fluorescent microscopy images to obtain quantitative information is not a trivial task due to several reasons: (1) not all cells will express GFP; (2) fluorescence seen in images can vary over time due to fluctuations occurring during the measurement process as well as other cellular functions; (3) some of the fluorescence seen in the images may be an artifact of the image. Image analysis algorithms are required in order to address these points. Accordingly, developing algorithms for analyzing fluorescent microscopy images of GFP reporter cells is an important step for obtaining quantitative data of protein concentrations in signal transduction pathways. One typical task for image analysis is to determine fluorescent cell regions from fluorescent images. Individual cells should be further identified in the case that fluorescent cells have obvious phenotype heterogeneity. The rationale behind this is that information about phenotype heterogeneity among individual cells, i.e. the fluorescence

intensity distribution, plays an important role for the dynamics of the underlying signal transduction pathways (Efroni et al., 2007; Smits et al., 2005). Detection of individual cells from fluorescence microscopy images has attracted significant interest in the computational biology community. Rosenfeld et al. (2005) and Spetsieris et al. (2009) presented techniques for identifying individual E. coli cells from fluorescence microscopy images. Compared with bacteria such as E. coli, which are generally rod-shaped, eukaryotic cells such as H35 and HepG2 do not have a regular shape or an easily detectable boundary. Furthermore, fluorescent images of these cells are characterized by low contrast and a significant noise level. Unfortunately, interpreting the results derived from standard edge detection approaches (Bao et al., 2005; Canny, 1986; Chaudhuri et al., 1989; Geback and Koumoutsakos, 2009; Haralick, 1984; Hsiao et al., 2006; Lia et al., 2009; Marr and Hildreth, 1980; Meyer, 1979; Perona and Malik, 1990; Poon et al., 1999; Prewitt, 1970; Roberts, 1965; Sobel, 1978; Torre and Poggio, 1979) is non-trivial for these images; instead, an algorithm that specifically addresses the above described image properties needs to be developed for determining boundaries of individual fluorescent cells from fluorescence microscopy images of GFP reporter systems.

While the quantitative data of transcription factors is potentially accessed via the approach based on GFP reporter systems, the available data might not be enough to verify the models which have hundreds of uncertain parameters. Model simplification is a promising approach to address this. Many different model reduction techniques exist, which can for the most part be put into one of the following categories of techniques: 1) retaining input–output properties of the system by balancing (Hahn and Edgar, 2002;

Liebermeister et al., 2005; Moore, 1981; Skogestad and Postlethwaite, 1997; Sun and Hahn, 2006a); 2) lumping several state variables into new state variables (Dano et al., 2006; Dokoumetzidis and Aarons, 2009); 3) separating the reactions based on time scale involving a quasi-steady-state assumption (Kruger and Heinrich, 2004); 4) eliminating reactions and species from the model by solving a mixed-integer nonlinear programming problem (Androulakis, 2000; Edwards et al., 1998; Petzold and Zhu, 1999; Sirdeshpande et al., 2001; Taylor et al., 2008). While these techniques can be applied to a wide-variety of problems, they also have drawbacks associated with them. For example, models reduced by balancing lose the physical significance of the states while the solution of mixed-integer nonlinear programming problems can be computationally expensive. These observations form the motivation behind this work: to develop a procedure for deriving simplified signal transduction models that retain the most important verifiable relationships between concentrations of several different proteins and apply this procedure to a detailed model of a signal transduction pathway of interest.

1.5 Dissertation Outline

The structure of this dissertation is described as follows:

Section 2 presents background information which is required to appreciate the contribution of the remainder of the work. It includes IL-6 signal transduction modeling, sensitivity analysis, observability analysis, K-means clustering, principal component analysis (PCA), solution of inverse problems, and mathematical morphology analysis.

Section 3 presents a model simplification procedure for signal transduction pathway models. The presented approach focuses on developing a simplified model where the

physical interpretation of the important states and parameters of the original model can be retained and the components can be verified by experimental data. The technique makes use of sensitivity analysis, parameter clustering, as well as observability analysis. The methodology is then applied to an IL-6 signal transduction pathway model which included 65 components and 111 parameters. The performance and identifiability of the simplified IL-6 model are then evaluated.

Section 4 presents an integrated modeling and experimental approach for quantitatively determining transcription factor profiles from GFP reporter data. The technique consists of three steps: (1) creating data sets for green fluorescent reporter systems upon stimulation, (2) analyzing the fluorescence images to determine fluorescence intensity profiles using PCA and K-means clustering, and (3) computing the transcription factor concentration from the fluorescence intensity profiles by solving an inverse problem from a model that describes transcription, translation, and activation of green fluorescent proteins. This technique is applied to quantitatively characterize activation of the transcription factor NF-κB by the cytokine TNF-α. In addition, the quantitative NF-κB profiles obtained from this technique are used to develop a model for TNF-α signal transduction where the parameters are estimated from the obtained data. Quantitative data for transcription factors nuclear STAT3 and C/EBPβ is also obtained and then used for verifying the simplified IL-6 model derived in Section 3.

Section 5 presents techniques for quantifying the fluorescence intensity distribution of GFP reporter systems and for computing the distribution of transcription factor profiles from the fluorescence intensity distribution. Unlike the approach presented in

Section 4, this section focuses on calculating the fluorescence intensity distribution among individual fluorescent cells instead of calculating just the average fluorescence intensity over fluorescent cell regions. The presented approach consists of an algorithm for identifying individual fluorescent cells from fluorescent images, and an algorithm to compute the distribution of transcription factor profiles from the fluorescence intensity distribution by solving an inverse problem. The technique is applied to experimental data to derive the distribution of the NF-κB concentrations from fluorescent images of a NF-κB GFP reporter system.

Section 6 summarizes the presented results and provides some suggestions for future work beyond the work of this dissertation.

2. PRELIMINARIES

2.1 Model Describing IL-6 Signal Transduction

The starting point for the model of the IL-6 signal transduction pathway used in this work is the one presented by Singh et al., 2006, which is based upon the model structure proposed by Heinrich et al., 2003, where the dynamic model of Jak-STAT signaling is adopted from Yamada et al., 2003, and parts of the detailed kinetic model of Erk-C/EBPβ signaling proposed by Schoeberl et al., 2002, are also used.

This model can be represented by a set of nonlinear ordinary different equations in the form of Eq. (1.8). The model presented by Singh et al., 2006, is updated by removing the components where the receptor complex has either STAT3 and SOSC3 or SHP2 and SOCS3 bound to it (Huang et al., 2009a). The rationale for this change is that SOCS3 competes for the same binding site of the receptor complex with SHP2 or STAT3 (Fischer and Hilfiker-Kleiner, 2008). The Erk-C/EBPβ pathway in the model presented by Singh et al., 2006 is further extended from Erk-PP to the activation of the transcription factor C/EBPβ (Roth et al., 2001). The updated version of the model consists of 65 ordinary differential equations representing the concentration balances of the individual proteins and protein complexes, 111 parameters describing reaction constants, and one input given by the IL-6 concentration. The ODE model for this pathway is available on the website (Steven et al., 2010) .

2.2 Sensitivity Analysis

Sensitivity analysis is a widely-used tool to investigate the impact of changes of

some variables, e.g., parameters, on the outputs of the system. Sensitivity analysis techniques are generally categorized as local (Frank, 1978; Tomovic, 1972) or global (Cukier et al., 1973; McRae et al., 1982; Zi et al., 2005) techniques.

Local sensitivity analysis computes the sensitivity profile $\partial\mathbf{y}/\partial\mathbf{p}^{\mathrm{T}}$, where $\mathbf{y}$ is the output vector of length n_y, from Eq. (2.2) and (2.3) for a model consisting of Eq. (1.8) and (2.1).

$$\mathbf{y} = \mathbf{h}(\mathbf{x}, \mathbf{p}, \mathbf{u}) \tag{2.1}$$

$$\frac{d}{dt}\frac{\partial \mathbf{x}}{\partial \mathbf{p}^{\mathrm{T}}} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}^{\mathrm{T}}}\frac{\partial \mathbf{x}}{\partial \mathbf{p}^{\mathrm{T}}} + \frac{\partial \mathbf{f}}{\partial \mathbf{p}^{\mathrm{T}}} \tag{2.2}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{p}^{\mathrm{T}}} = \frac{\partial \mathbf{h}}{\partial \mathbf{x}^{\mathrm{T}}}\frac{\partial \mathbf{x}}{\partial \mathbf{p}^{\mathrm{T}}} + \frac{\partial \mathbf{h}}{\partial \mathbf{p}^{\mathrm{T}}} \tag{2.3}$$

The sensitivity vector, $\mathbf{s}_i$, is sampled from the sensitivity profile at times t_j, $j = 1, 2, \ldots, n_t$, and then normalized by the nominal values of the output and the corresponding parameter, i.e., y_i^0 and p_i^0.

$$\mathbf{s}_i = \left[\frac{p_i^0}{y_1^0}\frac{\partial y_1(t_1)}{\partial p_i} \quad \cdots \quad \frac{p_i^0}{y_1^0}\frac{\partial y_1(t_{n_t})}{\partial p_i} \quad \cdots \quad \frac{p_i^0}{y_{n_y}^0}\frac{\partial y_{n_y}(t_1)}{\partial p_i} \quad \cdots \quad \frac{p_i^0}{y_{n_y}^0}\frac{\partial y_{n_y}(t_{n_t})}{\partial p_i} \right]^{\mathrm{T}} \tag{2.4}$$

One popular method to evaluate the effect of the parameter p_i on the output $\mathbf{y}$ is shown in Eq. (2.5) where the norm of the sensitivity vector is used as a measure of the impact of p_i on $\mathbf{y}$:

$$\|\mathbf{s}_i\| = \sqrt{\sum_k \sum_j (p_i^0 / y_k^0 \times \partial y_k(t_j) / \partial p_i)^2} \tag{2.5}$$

2.3 Observability Analysis of Nonlinear Systems

Observability represents the property of a system that allows the reconstruction of state variables from the given outputs (Brockett, 1970). While no readily available global observability analysis procedure exists for nonlinear systems, it is possible to approximate observability of a nonlinear system over a region in state space by using the observability covariance matrix (Hahn et al., 2003) with the following steps: firstly, perturbation directions such as plus and minus, and possible perturbation amplitudes are defined for each state to represent all the possible values of the state over an operation region; secondly, each state is taken as the only output of the system and observability covariance matrix is calculated by Eq. (2.6) such that each element of the matrix represents the correlation between the outputs for the two corresponding states that are perturbed according to the defined perturbation directions and amplitudes. Observability covariance matrix can be regarded as the sum of the variance-covariance matrix of the selected output, corresponding to different initial conditions, and over time (Hahn and Edgar, 2001). Observability covariance matrix can then show how well the information of states can be inferred with the measured outputs. Large perturbations are taken for the calculation of observability covariance matrix to capture the nonlinearity of the system within the region of operation. In this work, perturbations are regarded as variations of concentrations of proteins involved in signal transduction pathways.

$$\mathbf{W_O} = \int_0^\infty \mathbf{T}_l \Psi^{lq}(t) \mathbf{T}_l^{\mathrm{T}} dt \tag{2.6}$$

$$\mathbf{T}^{n_x} = \left\{ \mathbf{T}_1, \cdots, \mathbf{T}_r; \quad \mathbf{T}_i \in R^{n_x \times n_x}, \quad \mathbf{T}_i^{\mathrm{T}} \mathbf{T}_i = \mathbf{I}, \quad i = 1, \cdots, r \right\} \tag{2.7}$$

$$\mathbf{g} = \{c_1, \cdots, c_s; \quad c_i \in R, \quad c_i > 0, \quad i = 1, \cdots, s\} \tag{2.8}$$

$$\mathbf{E}^{n_x} = \{\mathbf{e}_1, \cdots, \mathbf{e}_{n_x}; \quad \text{standard unit vectors in } R^{n_x}\} \tag{2.9}$$

where r is the number of matrices for the perturbation directions, s is the number of different perturbation sizes for each direction, and $\mathbf{\Psi}^{lq}(t) \in R^{n_x \times n_x}$ corresponds to $\Psi_{ij}^{lq}(t) = (\mathbf{y}^{ilq}(t) - \mathbf{y}_{ss})^{\mathrm{T}} (\mathbf{y}^{jlq}(t) - \mathbf{y}_{ss})$ where $\mathbf{y}_{ss}$ is the steady state value of the output for the corresponding perturbation and $\mathbf{y}^{ilq}(t)$ is the output of the system with the initial condition $\mathbf{x}(0) = c_q \mathbf{T}_l \mathbf{e}_i + \mathbf{x}_{ss}$. l, equal to 1, 2, ⋯, r, is the index of perturbation direction while q, whose values range from 1 to s, is the index of perturbation size. Details for choosing appropriate values for the parameters can be found in the literature (Hahn and Edgar, 2002). A MATLAB program for computing the observability covariance matrix is available online (Sun and Hahn, 2006b).

In order to make comparisons between different degrees of observability of a system, it is necessary to introduce observability measures. One such measure is the largest eigenvalue of the observability covariance matrix:

$$\mu = \lambda_{\max}(\mathbf{W_O}) \tag{2.10}$$

The observability measure μ can be computed for different potential measurements and a comparison of the values can be used to determine which measurements result in the most information that can be obtained about the system. A similar concept has been used in the field of sensor network design via observability analysis (Brewer et al., 2007; Singh and Hahn, 2005; Singh and Hahn, 2006). In this work, observability analysis is used to determine which proteins of a signaling pathway would allow to obtain

information about the signal transduction dynamics. It is recommended that the identified proteins be retained in the model.

2.4 K-means Clustering

Standard K-means clustering is a method for identifying patterns in data and for dividing data into k disjoint clusters (Kaufman and Rousseeuw, 1990). The principle of K-means clustering is to minimize the objective function shown in Eq. (2.11) by determining centroids for each of the *k* clusters:

$$\min_{\mu} F = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - x_{\mu,i} \right\|^2 \tag{2.11}$$

where S_i, i=1, 2, …, k, represents all points belonging to the i-th cluster, $x_{\mu,i}$ is the centroid of all the points $x_j \in S_i$, and x_μ is the collection of all the centroids. $x_{\mu,i}$ is calculated by Eq. (2.12).

$$x_{\mu,i} = \frac{\sum_{x_j \in S_i} x_j}{N_i} \tag{2.12}$$

where N_i is the total number of the data points in cluster S_i.

The procedure to perform K-means clustering consists of the following steps:

1) The initial centroids $x_{\mu,i}$, i=1, 2,… , k, for the k clusters are assigned or randomly sampled from the data points;
2) Each data point x_j is assigned to a cluster m. This decision is made by determining the smallest value for $\left\| x_j - x_{\mu,m} \right\|^2$ among all possible ones $\left\| x_j - x_{\mu,i} \right\|^2$, i=1, 2, …, k.

3) The function F from Eq. (2.11) is evaluated by computing the sum of the distances for all data points as well as for all clusters.
4) Eq. (2.12) is used to update the centroid of each cluster by averaging the data points of the corresponding cluster;
5) Steps 2) through 4) are repeated iteratively until the relative change in the objective function F between iterations is less than a certain threshold. The iterative refinement procedure is known as Lloyd's algorithm (Lloyd 1982; Sabin and Gray, 1986).

The key point for K-means clustering is the selection of the initial centroids for the k clusters. A proper choice for the initial centroids will make the clustering algorithm converge faster to the optimal solution.

2.5 Principal Component Analysis (PCA)

Principal component analysis (Hotelling, 1933) is a well-established technique for identifying multivariable patterns in data. A data matrix $\mathbf{X}$ can be composed as follows using PCA:

$$\mathbf{X}=\mathbf{M_s}\mathbf{M_L}^{T}+\mathbf{Er} \tag{2.13}$$

where $\mathbf{M_s}$ is the score matrix, $\mathbf{M_L}$ is the loading matrix, and $\mathbf{Er}$ is the residual between the actual data and the reconstruction by PCA. The columns of $\mathbf{M_L}$ represent principle components of the data matrix, while the columns of $\mathbf{M_s}$ are the projections of the data matrix onto the principle components (Jackson, 2003).

The motivation for using PCA for image analysis comes from the work presented in (Bharati and Macgregor, 1998; Geladi and Grahn, 1996), which shows that clusters in a

score plot from PCA are associated with features of an image. Furthermore, combining K-means clustering and PCA has been widely studied for clustering (Ding and He, 2004).

2.6 Solution of Inverse Problems

For the type of nonlinear systems given by Eq. (1.8) and (2.1), one type of inverse problem determines the input u from the measurements, $\hat{y}$, of the output y (Feldmann et al., 1998). This computation is usually performed by minimizing an objective function

$$F = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{2.14}$$

consisting of the sum of the squares over N measurements of the differences between the experimental data $\hat{\mathbf{y}}$ and the predicted output $\mathbf{y}$ computed from the model. The solution of this inverse problem is the trajectory of the input u.

There are two general categories of approaches for solving inverse problems: the non-parametric and the parametric approach. One common non-parametric approach deals with determining u with derivatives of certain orders of y (Piazzi and Visioli, 2000; Puebla and Alvarez-Ramirez, 2001). This technique may result in problems associated with numerical differentiation of data, as differentiation accentuates high-frequency noise and measurement errors (Benyon, 1979). Another non-parametric approach views the system inversion as a controller design problem (Dudley, 1985; Markusson, 2002; Sun and Tsao, 1999; Vollmera and Raisch, 2006) as shown in Fig. 6. The goal of this formulation is to determine the controller output u that minimizes the difference between the measurement $\hat{y}$ and the predicted output y.

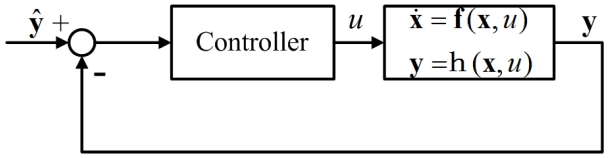


Figure 6. The scheme for tracking-controller design to minimize the difference between the experimental data $\hat{y}$ and the predicted output y computed from the model.

Unlike the non-parametric approach, the parametric approach solves inverse problems by parameterizing the input

$$u = g(\varphi_1, \varphi_2, \cdots, \varphi_m, t) \tag{2.15}$$

where $\varphi_i, i = 1,2,\cdots,m$, are parameters to be determined and g describes the effect of these parameters on u (Dudley, 1985). The parametric approach transforms the inverse problem to an optimization problem where the parameters $\varphi_i, i = 1,2,\cdots,m$, are estimated by minimizing the objective function shown in Eq. (2.14).

$$\begin{aligned} &\min_{\varphi_1,\varphi_2,\cdots,\varphi_m} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \\ &subject \quad to \\ &\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, \mathbf{p}, g(\varphi_1, \varphi_2, \cdots, \varphi_m, t)) \\ &\mathbf{y} = \mathbf{h}(\mathbf{x}, \mathbf{p}, g(\varphi_1, \varphi_2, \cdots, \varphi_m, t)) \\ &\mathrm{y_i} = y(t_i) \end{aligned} \tag{2.16}$$

One special form of parameterizing u is given by

$$u = \varphi_i, \quad \varphi_i = u(t_i) \quad for \quad t_i < t \le t_{i+1} \tag{2.17}$$

where the inverse problem of determining u at each point in time simplifies to computing u only at the sampling points and keeping it constant between the sampling

points.

Computationally solving inverse problem can involve the solution of a constrained nonlinear programming problem as shown in Eq. (2.16) or viewing the system inversion as a controller design problem as shown in Fig. 6. While other approaches exist, they are not further described in this section as the focus of the presented approach is on the parametric approach and determining a solution via solution of an optimization problem.

2.7 Mathematical Morphology Analysis

Mathematical morphology (Serra, 1982) is a well-established theory for processing binary and grayscale images which has been extensively applied in the area of image noise reduction (Peters, 1995) and pattern recognition (Hsiao et al., 2006; Lia et al., 2009; Yamamoto et al., 1996). The basic idea behind mathematical morphology is to use predefined structures, such as a matrix containing only elements with a value of 0 or 1, to probe or modify the pixels of images. Morphological image processing is based on a set of morphological operations, such as dilation, erosion, opening and closing (Haralick et al., 1987). Algorithms for morphological image processing are available in many software packages, e.g., the Image Processing Toolbox of MATLAB.

In this dissertation, mathematical morphology is used in Section 5 for image pre-processing, such as removing isolated noisy pixels from images and obtaining an initial estimate of the distance between different fluorescent cells.

3. MODEL SIMPLIFICATION PROCEDURE FOR SIGNAL TRANSDUCTION PATHWAY MODELS: AN APPLICATION TO IL-6 SIGNALING

3.1 Overview

This section presents a procedure for deriving simplified signal transduction models that retain the most important verifiable relationships between concentrations of several different proteins and apply this procedure to a detailed model of a signal transduction pathway of interest. The presented work uses sensitivity analysis where several important proteins, for which measurement data can be generated, are chosen as the outputs of the model. Model parameters are clustered based upon their sensitivity profiles for each output. Parameters that belong to the same cluster indicate that the mechanisms involving these signaling intermediates can be simplified. Representative state variables are then chosen for the reactions associated with each cluster of parameters via an observability analysis measure. This norm determines which proteins in the signaling pathway should be measured to maximize the information that can be extracted about the dynamics of proteins of the signal transduction pathway. A simplified model is then constructed based upon the selected state variables and the parameter clustering results.

The presented procedure will be illustrated by applying it to the IL-6 signal transduction pathway. The reason for choosing this application is that IL-6 is one of the most important mediators for inflammation, several mathematical models of the IL-6 signaling pathways have been published (Huang et al., 2007; Singh et al., 2006), and that

it is difficult to validate these complex models as only a limited amount of quantitative data is available.

This section is structured as follows: Subsection 3.2 describes a procedure that guides the development of simplified signal transduction models. It is important to point out that this procedure represents more of a guideline than a technique that can as a whole be coded as a computer program. Subsection 3.3 uses the described procedure to derive a simplified model of the IL-6 signal transduction pathway. The derivation of the simplified model is an important contribution of this work as is evident from the complexity of the model and the detail provided for the procedure. Summary remarks are given in Section 3.4.

3.2 Derivation of Simplified Models for Signal Transduction

This section describes a new technique, consisting of a set of guidelines, which can be used for deriving simplified models of signal transduction pathways from complex models such that the model can be validated with limited experimental data. The methodology consists of four steps: (1) cluster the parameters of the model according to their sensitivity vectors; (2) select at least one representative state variable for the reactions associated with each cluster of parameters via observability analysis; (3) estimate parameters of the simplified model, the structure of which is determined by Steps (1) and (2); and (4) validate the simplified model with available experimental data.

3.2.1 Parameter Clustering via Sensitivity Analysis

This work uses an approach where the entire sensitivity profile, i.e., the information given by Eq. (2.4), is used for the analysis rather than just a norm of the sensitivity

profile, e.g., Eq. (2.5). The importance of this point is illustrated in Fig. 7 where two sensitivity profiles are plotted. Both profiles show very different behavior, yet they return the same value if the 2-norm is applied to the sensitivity vectors.

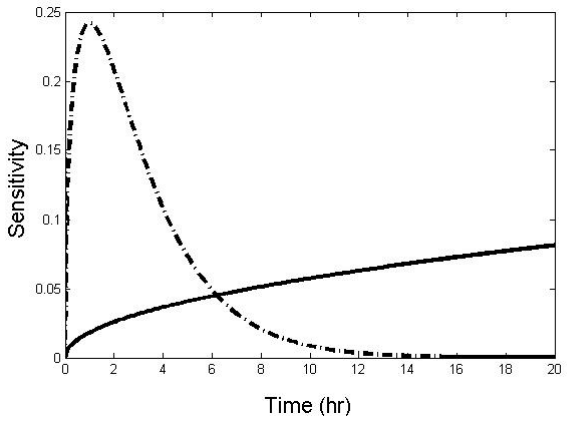


Figure 7. Two different sensitivity profiles. one has a peak while the other monotonically increases. However, they result in the same 2-norm value.

One option to take information about the entire sensitivity profile into account is to compute the angle between the sensitivity vectors corresponding to these profiles:

$$\cos\phi_{ik} = \frac{\left|\mathbf{s}_i^T \mathbf{s}_k\right|}{\left\|\mathbf{s}_i\right\|_2 \left\|\mathbf{s}_k\right\|_2} \tag{3.1}$$

where $\mathbf{s}_i$ and $\mathbf{s}_k$ are sensitivity vectors associated with the parameters p_i and p_k, respectively, and the cosine of the angle between the sensitivity vectors, $\cos\phi_{ik}$, defines a similarity measure. If the similarity measure is equal to unity then the effects that the two parameters have on the output are perfectly correlated.

The procedure for parameters clustering via sensitivity analysis consists of the following four steps:

1) For each reaction parameter of the model that is in the form of Eq. (1.8) and (2.1), the corresponding sensitivity vector is calculated by Eq. (2.2) ~ (2.4).
2) Parameters whose sensitivity vectors have a small length (e.g., less than 1% of the largest one) are removed as these parameters have a negligible effect on the output.
3) Based upon the similarity measure given by Eq. (3.1), the reaction parameters that have only a small angle between their sensitivity vectors are grouped via agglomerative hierarchical clustering. The outcome of this clustering procedure is a hierarchical tree that can be represented in a dendrogram.
4) A threshold for the hierarchical tree is selected to determine the number of groups of parameters. The reaction parameters of the signal transduction pathway are then clustered into the determined number of groups. Reaction parameters within the same group have a highly correlated effect on the output. The outcome of this procedure is that the parameters are clustered into distinct groups where changes of a parameter in a group can be compensated for by changes of other parameters within this group. Due to this correlation, it is only possible to estimate one parameter per group, which serves as an indicator that the model can be simplified without losing accuracy.

3.2.2 Determination of the Proteins Retained in the Simplified Model via Observability Analysis

The sensitivity analysis described in Subsection 3.2.1 investigates which reaction parameters have highly correlated effects. This information can serve as one indicator for

possible model simplification. Since the reactions associated with the parameters in the same cluster have highly correlated effects on the measurements, one representative reaction for each cluster of parameters should be chosen. While the number of reactions of the signaling pathway can be reduced using this procedure, it is still required to also reduce the number of state variables associated with these reactions. This section focuses on the approach used for selecting representative state variables from the ones that appear in reaction terms where the parameters can be found in the same cluster. The selected state variables associated with different clusters of parameters are then connected via a simplified reaction network.

Analyzing the degree of observability of a system reveals how much information about other states can be reconstructed from measuring specific states. It is suggested in this work to perform observability analysis to determine which of the states that are associated with the parameters contained in a cluster should be retained in the reduced model. This procedure consists of the following steps:

1) Only the state variables involved in reactions associated with the clustered parameters are considered as the potential components retained in the reduced model. The reason for this is that the parameters retained for clustering are known to have an important effect, which is described by the length of the sensitivity vector. Thus, only the state variables associated with these parameters can capture the main characteristics of the system dynamics. Additionally, state variables are put into groups corresponding to the parameters that they are associated with via the reaction terms.

2) One state variable at a time is assumed to be the only measurement of the system. The observability covariance matrix is computed according to Eq. (2.6). The largest eigenvalue of the observability covariance matrix is used as a measure for the degree of observability of the system. Next, a different state variable is chosen as the only measurement of the system and the procedure is repeated until the observability resulting from individually measuring each state variable has been analyzed.
3) The values of the observability measure for each selected state variable are compared. The components retained in the reduced model are then chosen by the following rationale: a) at least one state is selected for each cluster; b) state variables with higher values of the observability measure have a higher priority to be chosen; c) if two states corresponding to the same cluster have similarly large values of the observability measure, then the state is retained if it is associated with reactions connecting different clusters where the reactants are associated with one cluster of parameters via reaction terms while the products are associated with another cluster of parameters; d) states that can be directly measured in practice should also be retained.

The structure of the reduced model is determined by linking the chosen state variables from different clusters with the reactions connecting these clusters or by lumping the reactions in the same cluster into one reaction associated with the representative state variable. It may be necessary to retain some states other than the ones determined by this analysis in order to satisfy conservation laws.

3.2.3 Model Parameter Estimation

Since the newly derived model will include only a small fraction of the components of the original model, it is required to re-estimate the model parameters. The original model can be used to generate dynamic data of several of the key components of the system for parameter estimation for different stimulation profiles. The data set should be split up into a training set and a testing set. Parameter estimation can be expressed as an optimization problem

$$\begin{aligned} &\min_{\mathbf{p}} \sum_{k=1}^{n_y} \sum_{j=1}^{n_t} (\hat{y}_k(t_j) - y_k(t_j))^2 \\ &\textit{subject to} \\ &\quad \dot{\mathbf{x}} = \mathbf{f}_{\mathrm{r}}(\mathbf{x}, \mathbf{p}, \mathbf{u}) \\ &\quad \mathbf{y} = \mathbf{h}_{\mathrm{r}}(\mathbf{x}, \mathbf{p}, \mathbf{u}) \end{aligned} \tag{3.2}$$

where $\hat{y}_k(t_j)$ is the output of the original model, $y_k(t_j)$ is the prediction of the simplified model with the estimated parameters $\mathbf{p}$, $\mathbf{f}_{\mathrm{r}}$ and $\mathbf{h}_{\mathrm{r}}$ are functions representing the reduced model where the subscript ‘r’ represents the meaning of being reduced, n_y is the number of outputs, and n_t is the sample number. Parameter estimation can be performed using standard nonlinear least squares optimization routines such as lsqnonlin from MATLAB.

3.2.4 Validation of the Simplified Model with Experimental Data

The procedure described in the previous subsections allows construction of a model of lower complexity that should adequately describe the dynamics of the outputs of the original model. However, this alone may be insufficient as the ultimate goal of the model is not only to approximate parts of a more detailed model, but to provide a good representation of the real system.

As quantitative concentration measurements of proteins involved in signal transduction are rare. A technique has been developed in this dissertation and is shown in Section 4 to obtain transcription factor concentrations, e.g., for nuclear STAT3 or NF-κB, by analyzing fluorescent microscopy images of a GFP-reporter system and solving an inverse problem from a model that links the fluorescence intensity to the transcription factor activity. In that technique, fluorescence intensity is quantified from fluorescent images and a model is developed to describe the dynamics exhibited by the components involved in the GFP-reporter due to the induction of transcription factors. The model describing GFP dynamics is integrated with the signal transduction pathway model here for parameter estimation. The derivation of g and the calculation of $\hat{I}$ in Eq. (3.3) are discussed in Section 4.

$$\begin{aligned} &\min_{\mathbf{p},R_m} \quad \sum_{j=1}^{n_t} (\hat{I}(t_j) - I(t_j))^2 \\ &subject \quad to \\ &\dot{\mathbf{x}} = \mathbf{f}_r(\mathbf{x},\mathbf{p},\mathbf{u}) \\ &C_{TF} = \mathbf{h}_r(\mathbf{x},\mathbf{p},\mathbf{u}) \\ &I = g(C_{TF}) \end{aligned} \tag{3.3}$$

where I is the average fluorescence intensity over all cells seen in the images, which is predicted by the model, $\hat{I}$ refers to the experimental data, C_{TF} is the concentration of transcription factor in the nucleus, g is the model linking C_{TF} to fluorescence intensity I.

In this integrated model, the output of the model for signal transduction pathway, i.e. the concentration of transcription factor C_{TF}, is the input of the model for the GFP-reporter system. Parameters of the model for signaling pathway can be estimated by

fitting the fluorescence intensity profile predicted by the integrated model, i.e. I in Eq.(3.3), to the fluorescence intensity data generated by GFP reporter systems.

3.3 Application of the Methodology to IL-6 Signal Transduction Model

The technique presented in Subsection 3.2 is applied to derive a simplified model for IL-6 signal transduction. IL-6 is one of the most important mediators for inflammation and many studies involving the IL-6 signaling pathway or models thereof have been presented in the literature (Fasshauer et al., 2004; Heinrich et al., 2003; Huang et al., 2007; Lang et al., 2003; Schoeberl et al., 2002; Singh et al., 2006; Yamada et al., 2003).

3.3.1 Parameter Clustering for the Reaction Parameters Involved in IL-6 Signaling

The approach presented in Subsection 3.2.1 is applied to cluster the reaction parameters of the IL-6 signaling model described in Subsection 2.1. The nuclear concentrations of the two transcription factors, STAT3 and C/EBPβ, are taken as the outputs of the model. A dendrogram is constructed once for STAT3 as the output and once for C/EBPβ as the output. As would be expected, the parameters associated with the pathway of which the transcription factor is measured for each of the two cases, STAT3 for Jak-STAT and C/EBPβ for Erk-C/EBPβ, have large sensitivity vectors. The dendrogram and clustering result for the parameters involved in the Jak-STAT pathway with relatively large sensitivity vectors are shown in Fig. 8 and Fig. 9, respectively. Four clusters of parameters are obtained by setting the threshold as 0.7 in the dendrogram. The reason for choosing this threshold here is that it represents a good trade-off between the model size of the simplified model and the mechanisms that will be retained. The same analysis approach that was used for measuring STAT3 has also been implemented

for C/EBPβ as a measurement. The results for measuring C/EBPβ are not shown here due to space constraints, however, they can be summarized as follows: (1) most of the parameters of the Jak-STAT pathway have small sensitivity vectors as the measurement is a component of the Erk-C/EBPβ pathway; (2) the components SHP2 and Erk-PP are involved in the reactions where the parameters have large sensitivity vectors and should be retained in a simplified model.

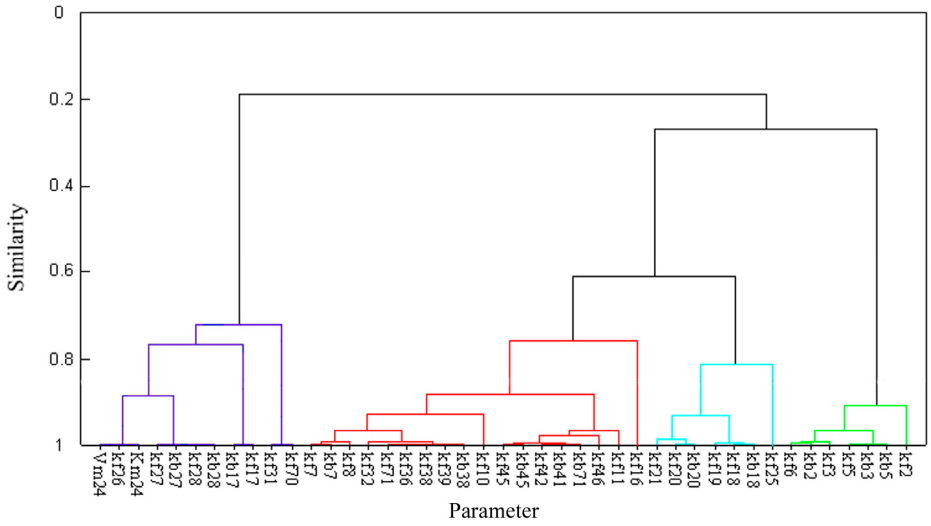


Figure 8. Dendrogram of the hierarchical tree of parameters when nuclear STAT3 is measured. Only the parameters involved in the Jak-STAT pathway with relatively large sensitivity vectors are shown here.

It can be concluded that the original model contains many parameters which cannot be determined from measuring the transcription factor activity, but also that the parameters that have a similar effect on the output are often associated with reactions involving a limited number of key components. One example of this is that several reaction parameters with closely correlated sensitivity vectors are associated with the formation and activation of STAT3. Similarly, the reaction parameters involved in signal

transduction through the cell membrane are also found in the same group.

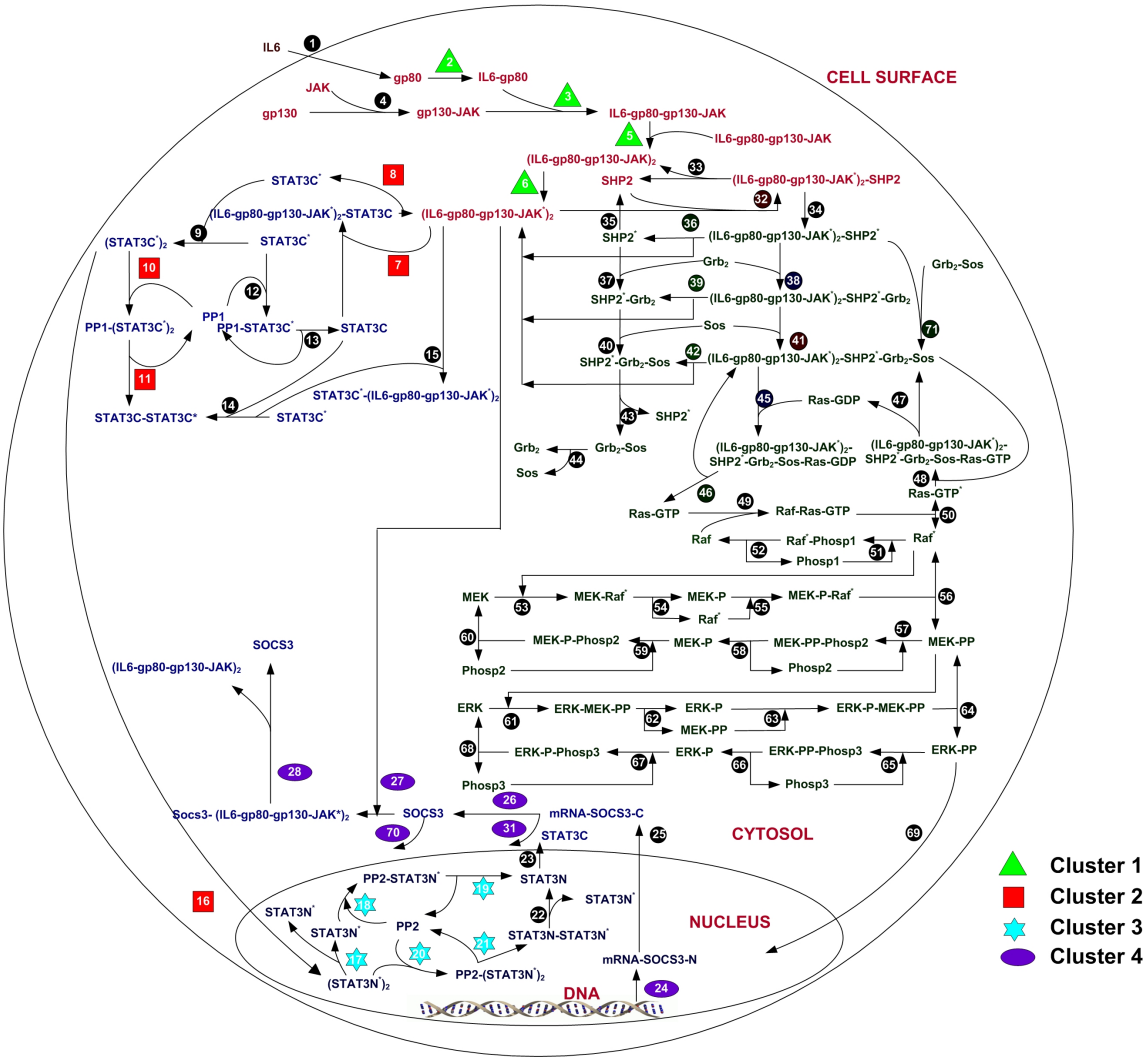


Figure 9. Sensitivity analysis results for the signal transduction model. Parameters have sensitivity vectors that are close to being collinear are shown in the same cluster. The reactions whose rate constants have relatively small sensitivity vectors are marked by black circles where the indicated numbers represent the indexes of the corresponding reactions in the signal transduction pathway.

3.3.2 Determining States of Simplified IL-6 Signaling Model

The approach presented in Section 3.2.2 is implemented to determine the states, i.e., protein concentrations, retained in the simplified model. The values of the observability

measure for the state variables involved in the Jak-STAT pathway are listed in Table 1 in decreasing order along with the cluster number that each state variable is associated with. It can be seen from Table 1 that $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$ has the highest value of the observability measure in Cluster 1. In addition, $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$ is associated with the reactions connecting Cluster 1 to Cluster 2 and Cluster 4: 1) In the reaction where $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$ combines with STAT3C to form $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$-STAT3C, $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$ is associated with Cluster 1 while the other two belong to Cluster 2 (please refer to Table 1); 2) In the reaction where $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$-SOCS3 is formed via binding of SOCS3 to $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$, both $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$-SOCS3 and SOCS3 are associated with Cluster 4. Therefore, the concentration of $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$ is chosen as the representative state variables of Cluster 1. In Cluster 2, the value of the observability measure of $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$-STAT3C is similar to that of $STAT3C^*$-$STAT3C^*$. As has been discussed above, $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$-STAT3C is directly involved in connecting Cluster 2 to Cluster 1, thus $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$-STAT3C is selected as the representative component of Cluster 2. The concentrations of $STAT3N^*$-$STAT3N^*$ and SOCS3 are chosen as the representative state variables for Cluster 3 and Cluster 4, respectively, because they have the largest value of the observability measures in their corresponding clusters and they are also involved in connecting their clusters to other clusters: $STAT3N^*$-$STAT3N^*$ results from STAT3C dimerizing and translocating to the nucleus, while it also serves as the transcription factor that results in the formation of SOCS3 after transcription/translation. These results are consistent with what is known

about IL-6 signal transduction (Heinrich et al., 2003; Huang et al., 2007; Singh et al., 2005): $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$ initiates signaling through both the Jak-STAT pathway and the Erk-C/EBPβ pathway, where $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$-STAT3C is associated with the reaction initiating signaling through the Jak-STAT pathway. $STAT3N^*$-$STAT3N^*$ is a transcription factor of the signal transduction pathway, while SOCS3 plays a key role as a signaling inhibitor of the Jak-STAT pathway.

Table 1. List of the 23 largest observability measures where each measure corresponds to only one component being measured.

Rank	Species Measured	Associated cluster	Observability Measure
1	SOCS3	Cluster 4	4727.7
2	mRNA-SOCS3C	Cluster 4	3973.9
3	mRNA-SOCS3N	Cluster 4	3190
4	$STAT3N^*$-$STAT3N^*$	Cluster 3	3091.7
5	$STAT3C^*$-$STAT3C^*$	Cluster 2	2836.7
6	$(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$	Cluster 1	2725.9
7	PP2-$STAT3N^*$-$STAT3N^*$	Cluster 3	2597.4
8	STAT3N-$STAT3N^*$	Cluster 3	2594.3
9	$(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$-STAT3C	Cluster 2	2539.1
10	STAT3C-$STAT3C^*$	Cluster 2	1924.2
11	STAT3N	Cluster 3	1445.5
12	$STAT3N^*$	Cluster 3	824.6
13	PP1-$STAT3C^*$	Cluster 2	729.1
14	$STAT3C^*$	Cluster 2	710.1
15	PP2-$STAT3N^*$	Cluster 3	620
16	gp130-JAK	Cluster 1	380.7
17	IL6-gp80	Cluster 1	360
18	$(IL6\text{-}gp80\text{-}gp130\text{-}JAK)_2$	Cluster 1	349.6
19	IL6-gp80-gp130-JAK	Cluster 1	298.4
20	$(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$-SOCS3	Cluster 4	290.9
21	PP2	Cluster 3	4.3
22	STAT3C	Cluster 2	1.5
23	PP1	Cluster 2	0.6

Once the representative components for each cluster of reactions have been chosen, several additional components need to be added such that the selected state variables of the signaling pathway can be linked without violating conservation laws. Based upon

this, STAT3C is selected to link (IL6-gp80-gp130-JAK)$^*_2$-STAT3C to (IL6-gp80-gp130-JAK)$^*_2$, while (IL6-gp80-gp130-JAK)$^*_2$-SOCS3 is chosen to link (IL6-gp80-gp130-JAK)$^*_2$ to SOCS3.

Table 2. Components of the simplified model for IL-6 signal transduction pathway.

Name	Species	Associated clusters (C)
x_1	(IL6-gp80-gp130-JAK)$^*_2$	Representative of C1
x_2	STAT3C	Connecting C1 and C2
x_3	(IL6-gp80-gp130-JAK)$^*_2$-STAT3C	Representative of C2
x_4	STAT3N*- STAT3N*	Output of Jak-STAT
x_5	SOCS3	Representative of C4
x_6	(IL6-gp80-gp130-JAK)$^*_2$-SOCS3	Connecting C4 and C1
x_7	SHP2	Representative of MAPK
x_8	(IL6-gp80-gp130-JAK)$^*_2$-SHP2-sum	Active form of SHP2
x_9	Erk-PP	Representative of MAPK
x_{10}	Erk	Inactive form of Erk-PP
x_{11}	(IL6-gp80-gp130-JAK)$_2$	Connecting C4, C1
x_{12}	C/EBPβi	Inactive form of C/EBPβn
x_{13}	C/EBPβn	Output of MAPK
u	IL-6	Input
R	Receptor	Connecting Input, C1

Selecting states of the simplified model for the Erk-C/EBPβ pathway is relatively simple, compared to the procedure described above for the Jak-STAT pathway as the Erk-C/EBPβ signaling pathway mainly consists of a series of sequential reactions. Since SHP2 and Erk-PP are involved in reactions having a key impact on C/EBPβ (see Subsection 3.3.1), these two components are retained in the simplified model. In addition, Erk and the complex involving SHP2 and (IL6-gp80-gp130-JAK)$^*_2$ are also retained in the simplified model to satisfy conversation laws. Specifically, Erk-pp is generated by phosphorylating Erk twice, and SHP2 exists in the cytoplasm in the form of SHP2 but may also be bound in a complex involving SHP2 and (IL6-gp80-gp130-

JAK)$^*_2$. Based on the above discussion, Table 2 shows the components of the simplified model.

Based upon the performed analysis, the reactions of the reduced model are given by Eq. (3.4) ~ (3.15). All the reactions are described by mass action kinetics, with the exception of Eq. (3.8) and Eq. (3.13), which are represented by Michaelis-Menten kinetics.

$$\mathrm{R} + \mathrm{IL6} \underset{p_2}{\overset{p_1}{\rightleftarrows}} (\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})^*_2 \tag{3.4}$$

$$(\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})^*_2 + \mathrm{STAT3C} \underset{p_4}{\overset{p_3}{\rightleftarrows}} (\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})^*_2\text{-}\mathrm{STAT3C} \tag{3.5}$$

$$(\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})^*_2\text{-}\mathrm{STAT3C} \xrightarrow{p_5} (\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})^*_2 + 1/2\ \mathrm{STAT3N}^*\text{-}\mathrm{STAT3N}^* \tag{3.6}$$

$$\mathrm{STAT3N}^*\text{-}\mathrm{STAT3N}^* \xrightarrow{p_6} \mathrm{STAT3C} + \mathrm{STAT3C} \tag{3.7}$$

$$\mathrm{STAT3N}^*\text{-}\mathrm{STAT3N}^* \xrightarrow{p_7, p_8, p_9} \mathrm{SOCS3} \tag{3.8}$$

$$\mathrm{SOCS3} \xrightarrow{p_{10}} \text{degradation} \tag{3.9}$$

$$(\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})^*_2 + \mathrm{SOCS3} \underset{p_{12}}{\overset{p_{11}}{\rightleftarrows}} (\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})^*_2\text{-}\ \mathrm{SOCS3} \tag{3.10}$$

$$(\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})^*_2\text{-}\ \mathrm{SOCS3} \xrightarrow{p_{13}} (\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})_2 + \mathrm{SOCS3} \tag{3.11}$$

$$(\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})^*_2 + \mathrm{SHP2} \underset{p_{15}}{\overset{p_{14}}{\rightleftarrows}} (\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})^*_2\text{-}\ \mathrm{SHP2\text{-}sum} \tag{3.12}$$

$$(\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})^*_2\text{-}\ \mathrm{SHP2\text{-}sum} + \mathrm{ERK} \xrightarrow{p_{16}, p_{17}} (\mathrm{IL6\text{-}gp80\text{-}gp130\text{-}JAK})^*_2\text{-}\ \mathrm{SHP2\text{-}sum} + \mathrm{ERKPP} \tag{3.13}$$

$$\mathrm{ERKPP} \xrightarrow{p_{18}} \mathrm{ERK} \tag{3.14}$$

$$\mathrm{ERKPP} + 2\ \mathrm{C/EBP\beta i} \xrightarrow{p_{19}} \mathrm{ERKPP} + \mathrm{C/EBP\beta n} \tag{3.15}$$

The resulting model consists of 12 reaction and 19 reaction parameters. A summary of the resulting signal transduction pathway model is shown in Fig. 10.

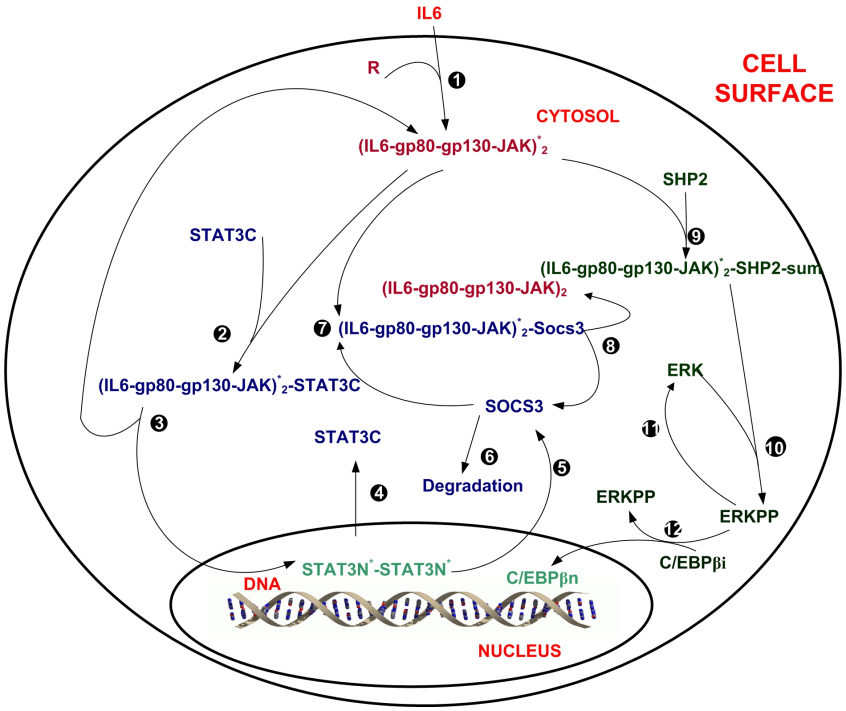


Figure 10. Overview of the simplified signal transduction model for the reactions shown in Eq. (3.4) ~ (3.15).

3.3.3 Parameter Estimation for the Simplified IL-6 Signaling Model and Performance Evaluation of the Simplified Model

The original model was used to generate dynamic data of several of the key components of the system ($STAT3N^*$-$STAT3N^*$, SOCS3, ErkPP, C/EBPβn). These data were used to obtain an initial estimate of the 19 parameters of the simplified model. The data set was split up into a training set and a testing set. As it is challenging to estimate the 19 reaction parameters all at once, several different data sets were initially created that correspond to different subsets of possible behaviors of the model: (1) initial estimates for all 19 parameters were computed from steady state data; (2) the parameters associated with Eq. (3.4) ~ (3.7) were estimated by blocking the pathway associated with SOCS3 and SHP2; (3) the parameters associated with Eq. (3.8) ~ (3.11) were estimated

by blocking the pathway associated with SHP2; (4) only the parameters associated with the Erk-C/EBPβ pathway were estimated. Steps (1)-(4) provide reasonable initial estimates for the parameters of the model. Using these initial estimates, all 19 parameters of the newly derived model were then estimated simultaneously in a final step.

The dynamics of the protein concentrations of the reduced model have been compared to the one from the original model and the predictions were found to be in good agreement. Refer to Fig. 11 for a comparison of four protein profiles.

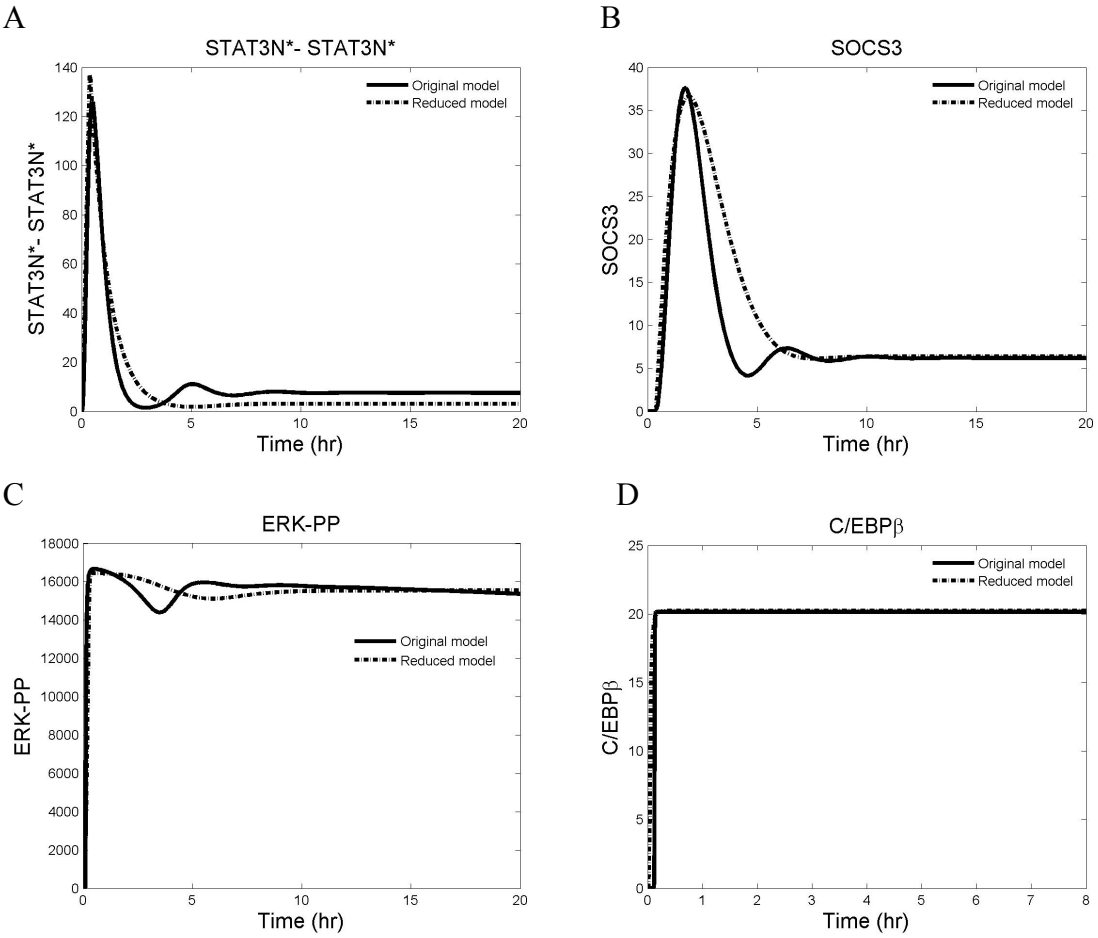


Figure 11. Comparison of model prediction for concentrations of nuclear STAT3, SOCS3, Erk-PP and C/EBPβn for the original and the simplified model.

In addition to performing comparisons by visual inspection, the relative errors (*Err*) have been computed for the outputs of the original and simplified model, i.e. $\hat{y}(t_i)$ and $y(t_i)$ shown in Fig. 11 according to the following formula

$$Err = \frac{\sum_i (\hat{y}(t_i) - y(t_i))^2}{\sum_i (\hat{y}(t_i))^2} \tag{3.16}$$

Table 3. Relative errors for the comparison results shown in Fig. 11.

State variables	*Err*
STAT3N*-STAT3N*	0.0771
SOCS3	0.1348
ERK-PP	0.0020
C/EBPβ	0.0027

As can be seen in Table 3, the largest relative error is 13.48%, which is associated with the profile of SOCS3. The other relative errors are below 10%. This relative error of the model is within the range of measurement errors that would be encountered for measuring concentrations of a component inside living cells.

In addition to fitting data generated by the original model, the simplified model's ability to predict the dynamic behavior of the states for different conditions is evaluated. Two different scenarios are investigated in this subsection: predicting profiles of the states for a cell where SHP2 phosphorylation is blocked, predicting profiles for a cell where the value of the input and the values of the initial concentrations of the proteins vary randomly in a 30% range around the normal values.

The results for the first scenario, i.e., blocking SHP2 phosphorylation resulting in no

signaling activity in the Erk-C/EBPβ pathway, are shown in Fig. 12, and the associated relative errors are given in Table 4. It can be seen that the reduced model adequately describes the dynamics of these proteins. This is especially important as a structural change has been made to the model, both the original one but also the simplified model, and no further re-estimation has taken place.

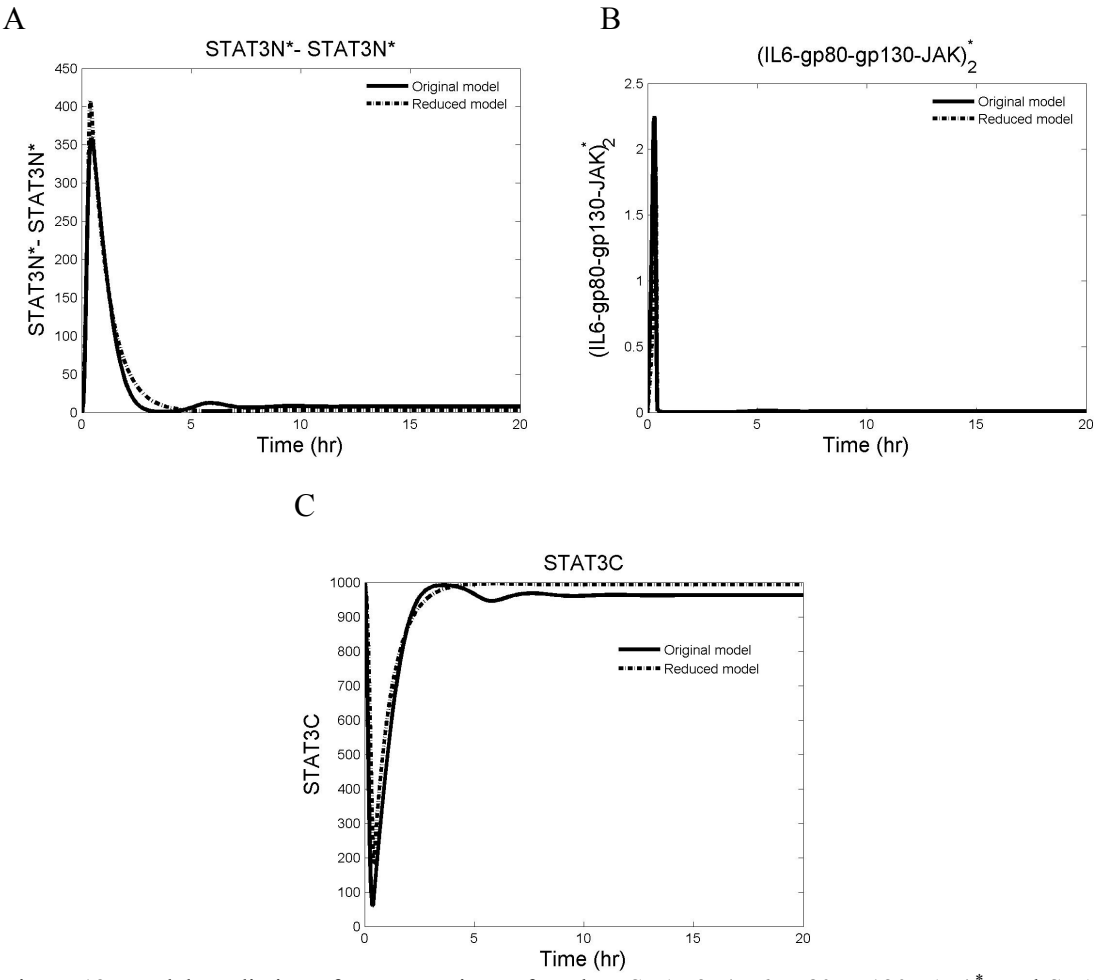


Figure 12. Model prediction of concentrations of nuclear STAT3, $(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$ and STAT3C in a cell with no signaling activity in the Erk-C/EBPβ pathway.

Table 4. Relative errors for the comparison results shown in Fig. 12.

State variables	*Err*
$STAT3N^*$-$STAT3N^*$	0.0181
$(IL6\text{-}gp80\text{-}gp130\text{-}JAK)^*_2$	0.2318
STAT3C	0.0226

The second scenario is investigated via Monte Carlo simulations consisting of the following steps:

1) The values of the input and the initial conditions for STAT3C, SHP2, Erk, and C/EBPβi, are sampled uniformly in the range of +70% ~ +130% of their corresponding normal values.

2) For each sampled set of values for the input and the initial conditions, the prediction from the simplified model is compared with that from the original model. The corresponding relative errors for nuclear STAT3 and C/EBPβ are calculated from Eq. (3.16).

3) Steps 1 and 2 are repeated 6000 times. This is defined as one set of Monte Carlo simulations. The reason for performing no more than 6000 simulations is that the population does not show statistically significant differences if more simulations are performed.

4) Step 1 through 3 are repeated 10 times to generate 10 data sets resulting from the Monte Carlo simulations.

5) The results of the relative errors for these 10 sets of Monte Carlo simulations are used to evaluate the prediction performance of the simplified model for the situation where a 30% uncertainty exits in the values of the input and the

initial conditions of the state variables.

Due to the space limitation, only the results for one set of Monte Carlo simulations are shown in Fig. 13. The mean value of *Err* for C/EBPβ is 0.0029 and the corresponding standard deviation is 0.0004, while the mean value of *Err* and the associated standard deviation for nuclear STAT3 are 0.1222 and 0.0515, respectively. It can be concluded that the simplified model can predict the dynamics of C/EBPβn very well even if the values of input and initial conditions for state variables vary around their nominal values.

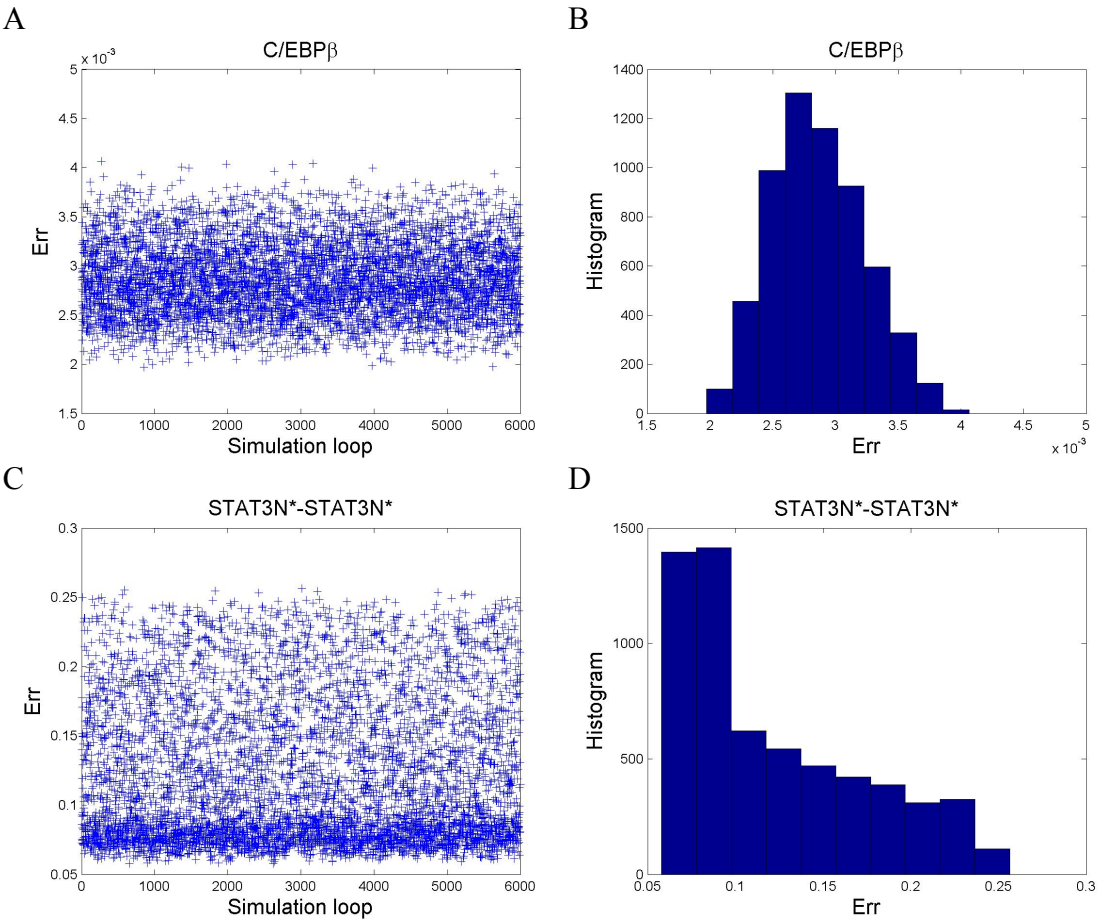


Figure 13. Relative errors of C/EBPβ and nuclear STAT3 for one set of Monte Carlo simulations.

One of the motivating factors for this investigation was that a model with a large number of parameters and only two measurements is likely overparameterized. Simplifying the model, and thereby reducing the number of parameters, should significantly improve identifiability. In order to investigate this aspect, the largest and the smallest singular values of the Fisher information matrix (FIM) of the simplified and the original model have been computed. The results are shown in Table 5. The condition number is defined as the ratio of the largest singular value to the smallest singular value in this case as FIM is a normal matrix. Based upon this, the condition number has been determined for the simplified model to be 4.87×10^4 and for the original model as 3.94×10^9. It can be concluded that reducing the number of parameters in the model has decreased the condition number of the Fisher information matrix by several orders of magnitude.

Table 5. Singular values and condition numbers of the FIM of the simplified and the original model.

Singular values and condition numbers	Simplified model	Original model
Largest singular value	52900	394000
Smallest singular value	1.09	0.0001
Condition number	4.87×10^4	3.94×10^9

3.4 Summary

It is common for signal transduction pathways that the involved proteins have been identified, however, little is known about the precise nature of the reaction mechanisms. A result of this is that many dynamic models of signal transduction pathways contain more detail than can be realistically verified given available experiment data. It is the main goal of this section to derive a model simplification procedure for signal

transduction pathways such that: (1) the model size is significantly reduced such that the model can be validated using available experimental data, and (2) the physical interpretation of the remaining states and parameters is retained. A model simplification procedure for signal transduction pathways is presented in this section. Sensitivity analysis is first performed to determine which parts of the model contain parameters that have highly correlated effects on the outputs of the system. These model parts can then be replaced by a simpler representation as it is not be possible to verify the values of all of the reaction parameters. Representative state variables are then chosen for each part of the model via quantification of the degree of observability of the state variables of the model for potential measurements. A new model structure can be derived based upon this analysis. The initial estimates of the parameters are generated from simulation data of the original model. In a final step, the parameters of the simplified model are re-estimated using available experimental data. The methodology has been applied to an IL-6 signal transduction pathway model. It was possible to reduce the original model which included 65 components and 111 parameters to a model with 13 components and 19 parameters. The reduced model was shown be able to reproduce the dynamics of important proteins with a reasonable degree of accuracy. The identifiability of the model has been improved significantly. A technique for obtaining quantitative data for transcription factors will be shown in Section 4. Based on the quantitative data obtained from this technique, the reduced model will then be verified. This will be shown in Section 4 either.

4. DERIVATION OF TRANSCRIPTION FACTOR PROFILES FROM FLUORESCENT REPORTER PROFILES

4.1 Overview

In this section, an integrated modeling and experimental strategy for deriving transcription factor activation rates from GFP-based fluorescent reporter systems is developed. The technique consists of three steps: (1) creating data sets for green fluorescent reporter systems upon stimulation, (2) analyzing the fluorescence images to determine fluorescence intensity profiles using PCA and K-means clustering, and (3) computing the transcription factor concentration from the fluorescence intensity profiles by inverting a model describing transcription, translation, and activation of green fluorescent proteins. This section only focuses on the last two step. The detail about the first step can be accessed in Huang et al., 2008. The quantitative data that is determined can be used to update models of signal transduction pathways. This is illustrated by first developing a model describing TNF-α signal transduction based upon the models presented by Lipniacki et al., 2004 and Rangamani and Sirovich, 2007, and then re-estimating model parameters from GFP reporter data for the activation of the transcription factor NF-κB by the cytokine TNF-α. The presented approach is not limited to NF-κB and can be used to determine the activation profile of any transcription factor as long as GFP reporter fluorescent profiles are available. This is illustrated by applying the presented technique to get quantitative data for validating the simplified IL-6 model obtained in Section 3.

4.2 Image Analysis Based on PCA and K-means Clustering

The series of images taken by fluorescence microscopy are analyzed to generate a time series of data representing the average fluorescence intensity of the cells in the images. In order to compute a fluorescence intensity profile, it is required to first determine the areas in the image representing cells where fluorescence can be seen. The procedure for determining these areas makes use of PCA and K-means clustering. A second step involves computing the average fluorescence intensity over these areas. The detailed steps involved in these procedures are described in the following. Each RGB image can be represented as a three-dimensional tensor.

$$\mathbf{M}_1 = \begin{bmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nm} \end{bmatrix}, \mathbf{M}_2 = \begin{bmatrix} g_{11} & \cdots & g_{1m} \\ \vdots & \ddots & \vdots \\ g_{n1} & \cdots & g_{nm} \end{bmatrix}, \mathbf{M}_3 = \begin{bmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nm} \end{bmatrix}, \mathbf{M} \in \Re^{n,m,3} \tag{4.1}$$

where the first two dimensions of $\mathbf{M}(i, j, k)$ refer to the position of a particular pixel on the image, i.e., the i-th row and j-th column, and the third dimension refers to the red ($k = 1$), green ($k = 2$), or blue ($k = 3$) value of the pixel. It is required to transform this three dimensional tensor, $\mathbf{M}$, to a two-dimensional matrix, $\mathbf{X}$ (Eq. (4.2)). Principal component analysis (Eq. (2.13)) is performed on $\mathbf{X}$ to determine pixels with similar brightness in the images. The columns of loading matrix $\mathbf{M_L}$ represent principle components of the image data matrix, while the columns of score matrix $\mathbf{M_s}$ are the projections of the image data matrix onto the principle components. An illustration of the data and the first principal component (PC1) is shown in Fig. 14. The projection of a point onto PC1 can be used as a measure for clustering the pixel brightness into different sets via K-means clustering. Fig. 15 illustrates the procedure of fluorescent cell searching based on K-mean clustering

and PCA. In an initial step PCA is used to divide the pixels of the image into two clusters based upon their projection onto PC1. K-means clustering iteratively updates the pixels and centroids of the two clusters until the sum of distances from all the pixels in each cluster is minimized. The cluster with the larger variation is divided in a next step. The centroids of the two new clusters, which are determined by PCA, and the centroid of the un-divided cluster are used as the initial centroids of the three clusters for K-means clustering, which then sorts the pixels of the image belonging to one of the three clusters. This procedure can be repeated until any number of desired clusters is obtained. The clusters with higher fluorescence intensity are considered to represent the cells which show a significant level of fluorescence.

$$\mathbf{X} = \begin{bmatrix} r_{11} & g_{11} & b_{11} \\ \vdots & \vdots & \vdots \\ r_{1m} & g_{1m} & b_{1m} \\ \vdots & \vdots & \vdots \\ r_{n1} & g_{n1} & b_{n1} \\ \vdots & \vdots & \vdots \\ r_{nm} & g_{nm} & b_{nm} \end{bmatrix} \tag{4.2}$$

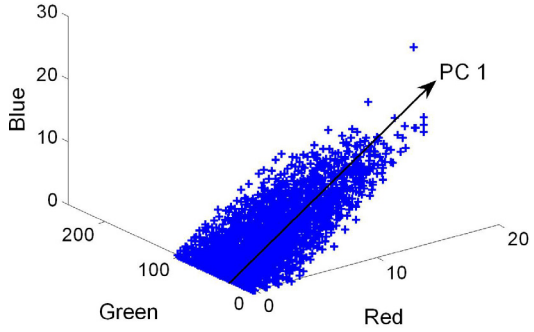


Figure 14. Principal component analysis of fluorescent images showing “green” as the principal component.

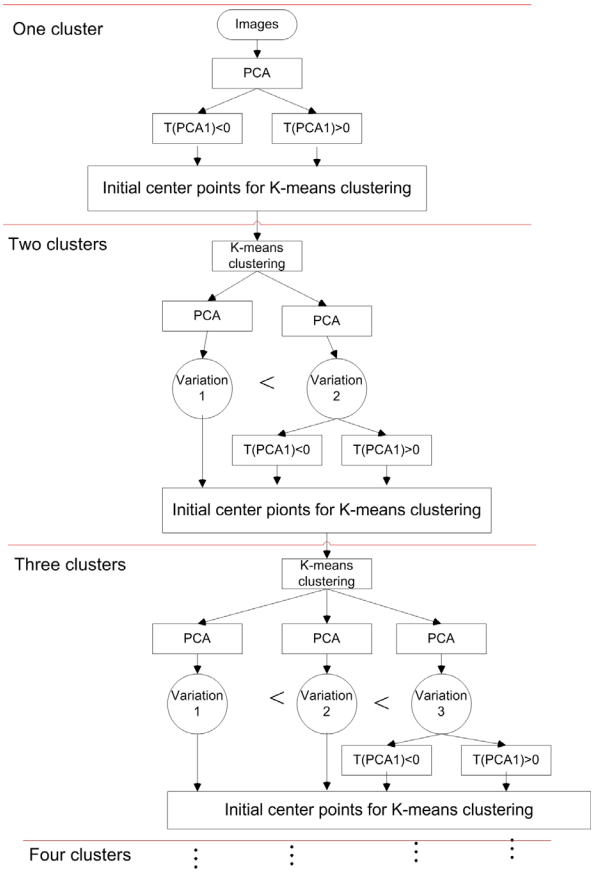


Figure 15. K-means clustering and PCA used for identifying cell regions in fluorescence images.

Once the cell region has been determined it is possible to compute the average fluorescence intensity by the following formula:

$$I = \left(\frac{\sum_{k=1}^{N_f} I_{f,k}}{N_f} - \frac{\sum_{k=1}^{N_b} I_{b,k}}{N_b}\right)_{stimulation} \tag{4.3}$$

$I_{f,k}$ refers the fluorescence intensity of the k_{th} pixel in a fluorescent cell region, $I_{b,k}$ refers the fluorescence intensity of the k_{th} pixel belonging to the background, N_f is the total number of pixels in the fluorescent cell region, N_b is the total number of pixels in the background. For a RGB image, the fluorescence intensity I is defined as the sum of the

values of red and green and blue of each pixel. The reason for subtracting the intensity of the pixels representing the background is to reduce measurement noise due to brightness variations. This procedure has to be repeated for each image taken at different points in time to generate a time series of data for the fluorescence intensity. An example of the outcome of this procedure can be seen in Fig. 16 where the first three clusters represent fluorescent cells while the pixels included in clusters 4 and 5 corresponds to the background. A graphic user interface (GUI) program for this image analysis technique is available on the website (Deimund et al., 2010).

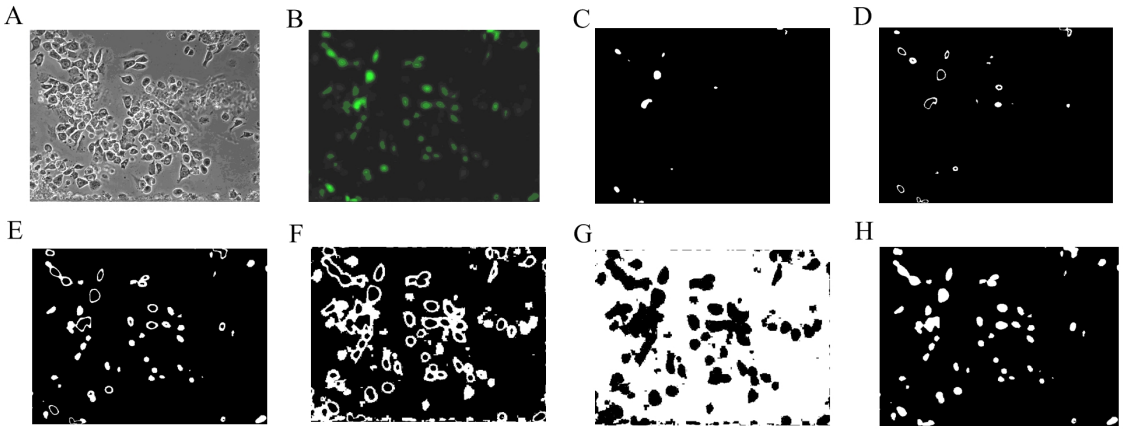


Figure 16. Results of the image analysis algorithm. (A) microscopy image, (B) Fluorescent regions detected by the image analysis procedure: (C) – (G) clusters 1 through 5 detected by the algorithm; white pixels refer to pixels included in a specific cluster, (H) cumulative results of clusters 1, 2, and 3; the white region in (H) is chosen as the region representing cells with GFP while the black pixels shown in (H) represent the background.

4.3 Derivation of a Model Describing GFP Dynamics

The dynamic model that used in this section is based upon the model published by Subramanian and Srienc, 1996, however, several modifications are made. Specifically, these changes are that:

1) the amount of DNA (C_{DNA}) remains constant in our work as the cells do not proliferate.
2) no growth dilution terms need to be included in the model for either the GFP m-RNA, m, balance (Eq. (4.4)), the non-fluorescent protein, n, balance (Eq. (4.5)), or the fluorescent protein, f, balance (Eq. (4.6)).
3) the transcription rate needs to be modified so that it depends on the amount of activated transcription factor present in the nucleus. This change results in the Monod kinetics shown in Eq. (4.4), $S_m \frac{C_{TF}}{C + C_{TF}} C_{\mathrm{DNA}}$, C_{TF} is the concentration of the transcription factor, replacing the original term which was solely based upon the amount of m-RNA present. While it was sufficient for the original model to neglect the transcription factor concentration, this is not the case for the model developed here as the transcription factor concentration is a crucial element of signal transduction and is regulated inside the cell.

The resulting model is given by Equations (4.4)-(4.6)

$$\frac{dm}{dt} = S_m C_{\mathrm{DNA}} \frac{C_{TF}}{C + C_{TF}} - D_m m \tag{4.4}$$

$$\frac{dn}{dt} = S_n m - D_n n - S_f n \tag{4.5}$$

$$\frac{df}{dt} = S_f n - D_n f \tag{4.6}$$

where m is the mRNA concentration; n is the concentration of GFP; f is the concentration of activated GFP; S_m is a reaction constant describing the transcription

rate with a value of 373 1/hr; C_{DNA} describes the amount of DNA and has a value of 5 nM; D_m is a constant describing the mRNA degradation rate and is equal to 0.8 1/hr; S_n is a reaction constant for the translation rate with a value of 780 1/hr; D_n is a constant associated with the protein degradation rate and is equal to 0.5 1/hr; S_f is associated with the fluorophore formation rate and has a value of 0.347 1/hr; C is a constant and the procedure for estimation of C is described in Subsection 4.5. The initial conditions for this system are $m(0)=0$, $n(0)=0$, and $f(0)=0$.

Equations (4.4)-(4.6) describe the relationship between the concentration of the transcription factor and activated GFP, f. The experimental measurements consist of the fluorescence intensity, I, from the images which is directly proportional to the concentration of activated green fluorescent protein:

$$I = f / \Delta \tag{4.7}$$

where Δ is the ratio between activated GFP and computed fluorescence intensity. As I can be obtained from the fluorescence images that have been processed by the procedures described in the image analysis section, the dynamics of transcription factors can be computed by solving an inverse problem involving equations involving equations (4.4)-(4.7).

4.4 Solution of an Inverse Problem

4.4.1 Problem Formulation

It is the purpose of the presented work to determine the profile of $C_{TF}(t)$ from Eq. (4.4)-(4.7) given the fluorescence intensity $I(t)$ over the time horizon of an experiment. A preliminary about the solution of inverse problems has been shown in Subsection 2.6,

stating that the parametric approach transforms the inverse problem to an optimization problem involving Eq. (2.16). This optimization problem can be non-trivial to solve as it involves the set of nonlinear differential equation from (4.4)-(4.7) as constraints. However, it should be noted that Eq. (4.4)-(4.7) do not describe a general nonlinear system but a Hammerstein system (Henson and Seborg, 1996) as a static nonlinear term, $\frac{C_{TF}}{C+C_{TF}}$, involving the input is coupled with linear differential equations. Due to this it is possible to introduce a transformation:

$$u = \frac{C_{TF}}{C + C_{TF}} \tag{4.8}$$

which results in a set of linear differential equations for the model from Eq.(4.5)-(4.7) and Eq. (4.9)

$$\frac{dm}{dt} = S_m C_{\mathrm{DNA}} u - D_m m \tag{4.9}$$

The optimization problem from Eq. (2.16) can now make use of the model given by the linear differentials Eq.(4.5)-(4.7) and Eq. (4.9) instead of the nonlinear equation from Eq. (4.4)-(4.7). The concentration profile for each sampling point can be computed after the optimization problem has been solved via the following equation:

$$C_{TF}(t_i) = C\frac{u(t_i)}{1-u(t_i)} \tag{4.10}$$

4.4.2 Illustrative Example Highlighting Limitations of Not Using Regularization

Since the problem formulation has been presented, one could proceed to applying this formulation to available fluorescence intensity data to compute the transcription

factor profile. In order to do this the tracking controller formulation briefly mentioned in Subsection 2.6 is used. As the model is linear, a linear state-feedback controller can be designed that minimizes the sum of the squares between the experimental data and the model predictions (Lin and Olbrot, 1996).

A linear quadratic regulator (LQR) controller is designed which matches the predicted intensity profile I to the experimental data $\hat{I}$ for the experimental data collected for stimulation of liver cells with 10 ng/ml of TNF-α. Fig. 17A shows the experimental results for the fluorescence intensity. Fig. 17A also depicts the predicted fluorescence intensity for the computed transcription factor profile shown in Fig. 17B. It can be seen that the experimental data can be fitted very well by this approach, however, the transcription factor NF-κB profile seen in Fig. 17B contradicts semi-quantitative data from the literature (Hoffmann et al., 2002) as well as predictions made by available models describing this signal transduction pathway (Lipniacki at al., 2004; Rangamani and Sirovich, 2007).

The reason for the results shown in this subsection is that the fluorescence intensity profile, even though it represents very good data for these types of measurements, is quite noisy. Solving an inverse problem with noisy data can lead to a profile for the input that is more affected by the noise than by the actual data. Therefore, it can be concluded that one should not try to perfectly fit the data. Instead, it is more meaningful to choose a more restrictive parameterization of the transcription factor profile such that realistic profiles can be obtained. Several candidates for this parameterization will be introduced in the following subsection.

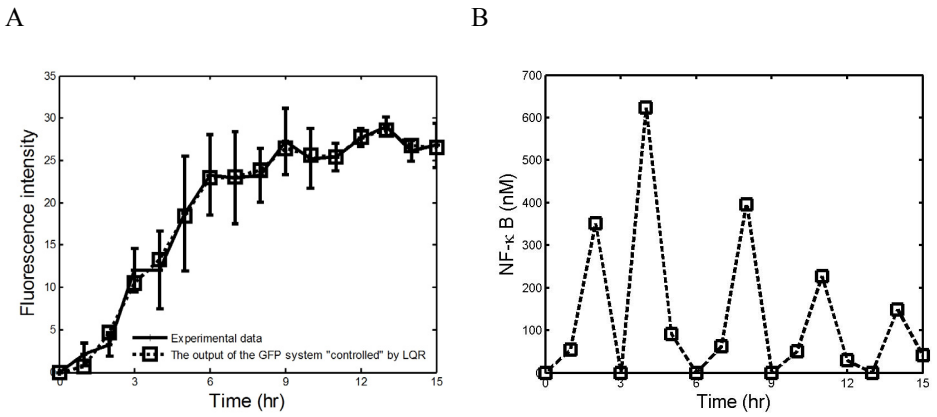


Figure 17. Results for using LQR controller approach to solve the inverse problem. (A) Fluorescence intensity profiles; (B) NF-κB profiles.

4.4.3 General Procedure for Computing Transcription Factor Profiles from Fluorescence Intensity Data

This subsection presents a procedure for solving an inverse problem that involves computing the transcription factor profiles from fluorescence intensity data. This procedure is robust to measurement noise as it is based upon parameterizing the transcription factor profiles such that the results are consistent with typical potential profiles. The technique is based on proposing several potential profiles and then analytically solving the set of differential equations given by Eq.(4.5)-(4.7) and Eq. (4.9). The solution of the inverse problem then only requires fitting of a few model parameters.

While this approach is straightforward, deriving the analytical solution can be rather tedious. Therefore, the details of all expressions for computing the parameters are provided in the Appendix as these equations form an important contribution of this work.

It is not unreasonable to assume that the vast majority of transcription factor dynamics can be described by the shape of one of the three profiles shown in Fig. 18.

Since these three profiles have fundamentally different shapes they will also result in different expressions, i.e., not just changes in parameters, for *u*. These three profiles for *u* can be described by, possibly delayed, step responses of a second-order under-damped system, of a second-order over-damped system, and of a lead-lag system augmented with a first-order filter.

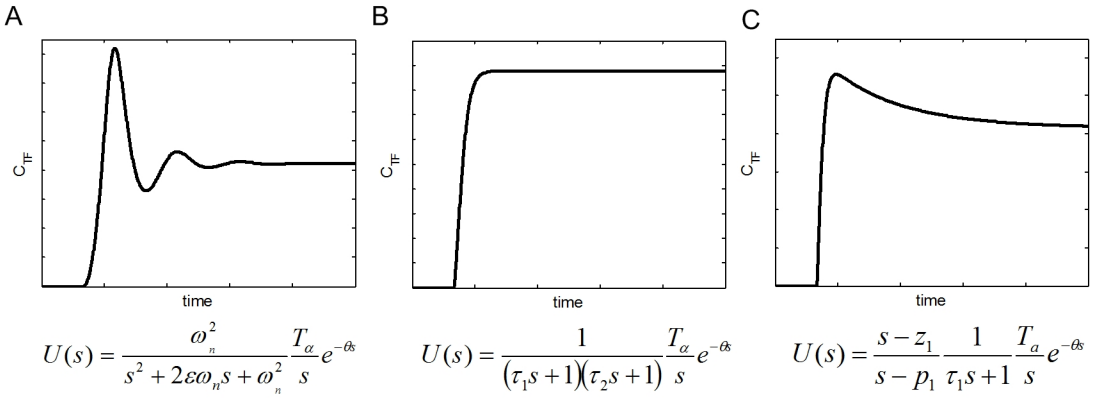


Figure 18. Potential profiles for *u* that will be investigated. Step response of (A) a second-order under-damped system with time delay, (B) a second-order over-damped system with time delay, and (C) a lead-lag system with time delay augmented with a first-order filter.

The rationale for selecting these profiles for the transcription factor dynamics is that profiles that exhibit (a) damped oscillations, (b) a continuous rise until they level off, or (c) a steep rise in concentration followed by a decline to a new steady state can be described. The extent to which this happens can vary from case to case and will result in different parameters for each expression of $u(t)$.

Table 6 shows the different expressions for $u(t)$ as well as their Laplace-transformed counterparts $U(s)$. It can be seen that each of these expressions includes several parameters which have to be determined from data as part of the overall procedure for

solving the inverse problem.

Table 6. Potential profiles for the input $u(t)$ shown in Fig. 18 in the time domain and frequency domain.

Potential profiles	Input profile in the time domain	Input profile in the frequency domain
Fig. 18.A	$u(t) = (T_\alpha - \frac{T_\alpha e^{-\varepsilon\omega_n(t-\theta)}}{\sqrt{1-\varepsilon^2}}\sin(\omega_n\sqrt{1-\varepsilon^2}(t-\theta)+\phi))$ $\cdot H(t-\theta)$	$U(s) = \frac{\omega_n^2}{s^2+2\varepsilon\omega_n s+\omega_n^2}$ $\cdot\frac{T_\alpha}{s}e^{-\theta s}$ where $0<\varepsilon<1$.
Fig. 18.B	$u(t) = (T_\alpha - \frac{T_\alpha\tau_1}{\tau_1-\tau_2}e^{-(t-\theta)/\tau_1}$ $-\frac{T_\alpha\tau_2}{\tau_2-\tau_1}e^{-(t-\theta)/\tau_2})H(t-\theta)$	$U(s) = \frac{1}{(\tau_1 s+1)(\tau_2 s+1)}$ $\cdot\frac{T_\alpha}{s}e^{-\theta s}$
Fig. 18.C	$u(t) = (\frac{z_1 T_a}{p_1} + \frac{(p_1-z_1)T_a}{p_1(\tau_1 p_1+1)}e^{p_1(t-\theta)}$ $-\frac{(\tau_1 z_1+1)T_a}{(\tau_1 p_1+1)}e^{-(t-\theta)/\tau_1})H(t-\theta)$	$U(s) = \frac{s-z_1}{s-p_1}\cdot\frac{1}{\tau_1 s+1}$ $\cdot\frac{T_a}{s}e^{-\theta s}$

Note: $H(t-\theta) = \begin{cases} 0, \text{ when } t<\theta \\ 1, \text{ when } t\geq\theta \end{cases}$ and $\phi = \arctan\frac{\sqrt{1-\varepsilon^2}}{\varepsilon}$

An analytical solution for the model describing transcription, translation, and activation of GFP, as given by Eq.(4.5)-(4.6) and Eq. (4.9), can be derived using Laplace transforms. Additionally, using Eq. (4.7) can lead to the following transfer function relating the fluorescence intensity $I(s)$ to the input of the system $U(s)$

$$I(s) = \frac{S_f}{s+D_n}\cdot\frac{S_n}{s+D_n+S_f}\cdot\frac{S_m C_{\text{DNA}}}{s+D_m}\cdot U(s) \tag{4.11}$$

where the different expressions for $U(s)$ from Table 6 can be substituted in Eq. (4.11).

For the case where $u(t)$ describes a damped oscillatory response, as the one seen in Fig. 18A, $I(s)$ is given by

$$I(s) = \frac{S_f}{s + D_n} \cdot \frac{S_n}{s + D_n + S_f} \cdot \frac{S_m C_{\mathrm{DNA}}}{s + D_m} \cdot \frac{\omega_n^2}{s^2 + 2\varepsilon\omega_n s + \omega_n^2} \cdot \frac{T_\alpha}{\Delta s} e^{-\theta s} \tag{4.12}$$

which in the time domain results in

$$\begin{aligned} I(t) = (A_1 + A_2 e^{-D_n(t-\theta)} + A_3 e^{-(D_n+S_f)(t-\theta)} + A_4 e^{-D_m(t-\theta)} \\ + A_5 e^{-\varepsilon\omega_n(t-\theta)} \sin(\omega_n \sqrt{1-\varepsilon^2}\,(t-\theta) + \varphi)) H(t-\theta) \end{aligned} \tag{4.13}$$

where the constants A_1 through A_5 are made up of the model parameters, which are fixed, as well as parameters that need to be determined by fitting $I(t)$ to the experimental data. The exact expression for these constants can be found in the Appendix A. This expression results in the following transcription factor dynamics

$$\begin{aligned} C_{TF}(t) = \frac{CT_\alpha\sqrt{1-\varepsilon^2} - CT_\alpha e^{-\varepsilon\omega_n(t-\theta)} \sin(\omega_n\sqrt{1-\varepsilon^2}\,(t-\theta) + \phi)}{\sqrt{1-\varepsilon^2} - (T_\alpha\sqrt{1-\varepsilon^2} - T_\alpha e^{-\varepsilon\omega_n(t-\theta)} \sin(\omega_n\sqrt{1-\varepsilon^2}\,(t-\theta) + \phi))} \\ \cdot H(t-\theta) \end{aligned} \tag{4.14}$$

where $\phi = \arctan\dfrac{\sqrt{1-\varepsilon^2}}{\varepsilon}$.

Similarly, if $u(t)$ refers to a monotone response, as shown in Fig. 18B, $I(s)$ is given by

$$I(s) = \frac{S_f}{s + D_n} \cdot \frac{S_n}{s + D_n + S_f} \cdot \frac{S_m C_{\mathrm{DNA}}}{s + D_m} \cdot \frac{1}{(\tau_1 s + 1)(\tau_2 s + 1)} \cdot \frac{T_\alpha}{\Delta s} e^{-\theta s} \tag{4.15}$$

which in the time domain results in

$$\begin{aligned} I(t) = (B_1 + B_2 e^{-D_n(t-\theta)} + B_3 e^{-(D_n+S_f)(t-\theta)} + B_4 e^{-D_m(t-\theta)} + B_5 e^{-(t-\theta)/\tau_1} + B_6 e^{-(t-\theta)/\tau_2}) \\ \cdot H(t-\theta) \end{aligned} \tag{4.16}$$

where the constants B_1 through B_6 can be found in the Appendix A. This expression

results in the following transcription factor dynamics

$$C_{TF}(t) = \frac{CT_a + Ca_1 e^{-(t-\theta)/\tau_1} + Ca_2 e^{-(t-\theta)/\tau_2}}{1 - T_a - a_1 e^{-(t-\theta)/\tau_1} - a_2 e^{-(t-\theta)/\tau_2}} H(t-\theta) \tag{4.17}$$

where $a_1 = -\frac{T_\alpha \tau_1}{\tau_1 - \tau_2}$, $a_2 = -\frac{T_\alpha \tau_2}{\tau_2 - \tau_1}$.

If $u(t)$ initially exhibits a sharp increase in value, as the one shown in Fig. 18C, then $I(s)$ can be described by

$$I(s) = \frac{S_f}{s + D_n} \cdot \frac{S_n}{s + D_n + S_f} \cdot \frac{S_m C_{\mathrm{DNA}}}{s + D_m} \cdot \frac{s - z_1}{s - p_1} \cdot \frac{1}{\tau_1 s + 1} \cdot \frac{T_\alpha}{\Delta s} e^{-\theta s} \tag{4.18}$$

which leads to

$$\begin{aligned} I(t) = {} & (C_1 + C_2 \exp(-D_n(t-\theta)) + C_3 \exp(-(D_n + S_f)(t-\theta)) + C_4 \exp(-D_m(t-\theta)) \\ & + C_5 \exp(p_1(t-\theta)) + C_6 \exp(-(t-\theta)/\tau_1))H(t-\theta) \end{aligned} \tag{4.19}$$

in the time domain where the constants C_1 through C_6 can be found in the Appendix A. This expression results in

$$C_{TF}(t) = \frac{(Cz_1 T_a(\tau_1 p_1 + 1) + CT_a(p_1 - z_1)e^{p_1(t-\theta)} - Cp_1 T_a(\tau_1 z_1 + 1)e^{-(t-\theta)/\tau_1})H(t-\theta)}{p_1(\tau_1 p_1 + 1) - z_1 T_a(\tau_1 p_1 + 1) - (T_a(p_1 - z_1)e^{p_1(t-\theta)} - p_1 T_a(\tau_1 z_1 + 1)e^{-(t-\theta)/\tau_1})H(t-\theta)} \tag{4.20}$$

for the transcription factor dynamics.

In summary, this subsection provided analytical solutions for the fluorescence intensity for three potential transcription factor profiles. Each of the three transcription factor profiles includes several parameters which also appear in the equation of the fluorescence intensity profiles. As a result of this, the optimization problem that was originally presented in Eq. (2.16) has been reduced to determining the parameters of the

input profiles shown in Table 6. This is a significantly simpler problem to solve as the parameters can be estimated using a nonlinear least squares optimization technique. The corresponding transcription factor profiles are then given by equations (4.14), (4.17), or (4.20), respectively. All that remains is to make a decision which of these three profiles fits the data best and discard the other two profiles.

The procedure for computing transcription factor profiles from fluorescence intensity data presented in this work consists of the following steps:

1) A data set of the fluorescence intensity $\hat{I}(t_i)$ is collected at N different points in time t_i, $i = 1...N$, by analyzing the average fluorescence intensity of cells using an image analysis procedure presented in Subsection 4.2.
2) Assuming that little is known about the transcription factor dynamics, it can be assumed that the transcription factor can follow any of the potential profiles investigated in this work. The optimization problem

$$\min_{\varphi_1,\varphi_2,\cdots,\varphi_m} \quad \sum_{i=1}^{N} (\hat{I}(t_i) - I_i)^2 \tag{4.21}$$

will be solved for each of the potential profiles given by equations (4.13), (4.16), and (4.19), where the parameters φ_i will be different for each equation. For example, for the intensity profile given by Eq. (4.13), the parameters that will be estimated are T_α, ε, ω_n, and θ, the parameters to be estimated for the profile from Eq. (4.16) are T_α, τ_1, τ_2, and θ, and the parameters for Eq. (4.19) are T_α, z_1, p_1, τ_1 and θ.

It should be pointed out that the parameters A_1 through A_5, B_1 through B_6, and C_1

through C_6 from equations (4.13), (4.16), and (4.19), respectively, can be directly computed for each of the three cases from these parameters via the equations provided in the Appendix A.

3) The intensity profile, among the three given by equations (4.13), (4.16), (4.19), that produces the smallest sum of squares error for Eq. (4.21), after the parameters have been fitted, is the one that best fits the experimental data.

4) If the intensity profile from Eq. (4.13) is the best one then the transcription factor concentration profile is given by Eq. (4.14) where the estimated values of the parameters T_α, ε, ω_n, and θ are used. Similarly, if the intensity profile from Eq. (4.16) produces the smallest sum of the squares error, then the transcription factor concentration from Eq. (4.17) with the values estimated for T_α, τ_1, τ_2, and θ provides the best description. Lastly, if Eq. (4.19) results in the best fit for the fluorescence intensity then the transcription factor profile is given by Eq. (4.20) with the estimated values for the parameters T_α, z_1, p_1, τ_1 and θ.

It should be pointed out that if one has a clear idea about the general shape of the transcription factor dynamics, then it is not necessary to solve the parameter estimation problem given by Eq. (4.21) for all three cases. Instead one can pick the profile that fits the knowledge about the system and just estimate the parameters for this particular profile.

The advantages that this procedure has over approaches that use a less restrictive parameterization are that the presented procedure is relatively insensitive to measurement error in addition to being computationally inexpensive. The first point is

particularly important as measurements of protein concentrations inside cells generally have a high noise level. However, this comes at the price of restricting the transcription factor profile to one of several possible candidates. That being said, the three potential dynamics cover a range of profiles that are generally assumed for transcription factors in response to continuous stimulation. If other profiles are found that are also deemed important than these, they can be added to the list of possible profiles for this technique.

4.5 Application to TNF-α Signal Transduction

Stimulating GFP-reporter liver cells with TNF-α will lead to the activation of transcription factor NF-κB (Hoffman et al., 2002; Lipniacki et al., 2004; Rangamani and Sirovich, 2007). It is of interest to know the transcription factor profile as NF-κB plays an important role in many cellular functions, however, directly measuring activated NF-κB in the nucleus is non-trivial. Using the procedure described in Subsection 4.2-4.4 allows to infer the NF-κB profile from fluorescence intensity data. It should be noted that while some qualitative knowledge about the NF-κB profile exists, e.g., it is assumed to exhibit damped oscillations (Hoffman et al., 2002; Lipniacki et al., 2004), nothing is known for certain as no quantitative measurement data for NF-κB are available. In this subsection, the procedures presented in Subsection 4.2-4.4 are applied to determine transcription factor NF-κB profiles from fluorescent images of a NF-κB GFP reporter system stimulated by TNF-α. A model for TNF-α signal transduction is first developed based on the models presented in Lipniacki et al., 2004 and Rangamani and Sirovich, 2007. Quantitative data is derived for NF-κB, which is then used to estimate the parameters of the developed model for TNF-α signal transduction. It should be noted

that the model describing the activation of NF-κB by TNF-α is not required for deriving NF-κB profiles from GFP profiles. However, use of the 1st principles model enables us to estimate model parameters using the data and thereby refine the model describing activation of NF-κB by TNF-α, so as to develop a systems level understanding of TNF-α signaling.

4.5.1 Model Development for TNF-α Signal Transduction

The model describing TNF-α mediated signal transduction is shown in Fig. 19 and the equations are given in Appendix B. This model is based upon the models described by Lipniacki et al., 2004 and Rangamani and Sirovich, 2007. The model from Lipniacki et al., 2004 is used to describe signal transduction from IKKn to NF-κB whereas the model from Rangamani and Sirovich's work is used to describe signal transduction from TNF-α to IKKn. The reason for combining these two models is that the model from Lipniacki et al.'s work does not describe signal transduction from TNF-α to IKKn, while the paper by Rangamani and Sirovich states that the signal transduction from IKKn to NF-κB as described in their model should be updated as it represents a simplification of what is currently known about the signal transduction pathway. In order to combine these two models the assumption that c-IAP in the reaction "Caspase-3*+c-IAP $\longrightarrow$ caspase-3*|c-IAP" from Rangamani and Sirovich's model can be replaced with cgent from Lipniacki et al.'s model. The rationale behind this assumption is that c-IAP and cgent are both involved in transcription of DNA.

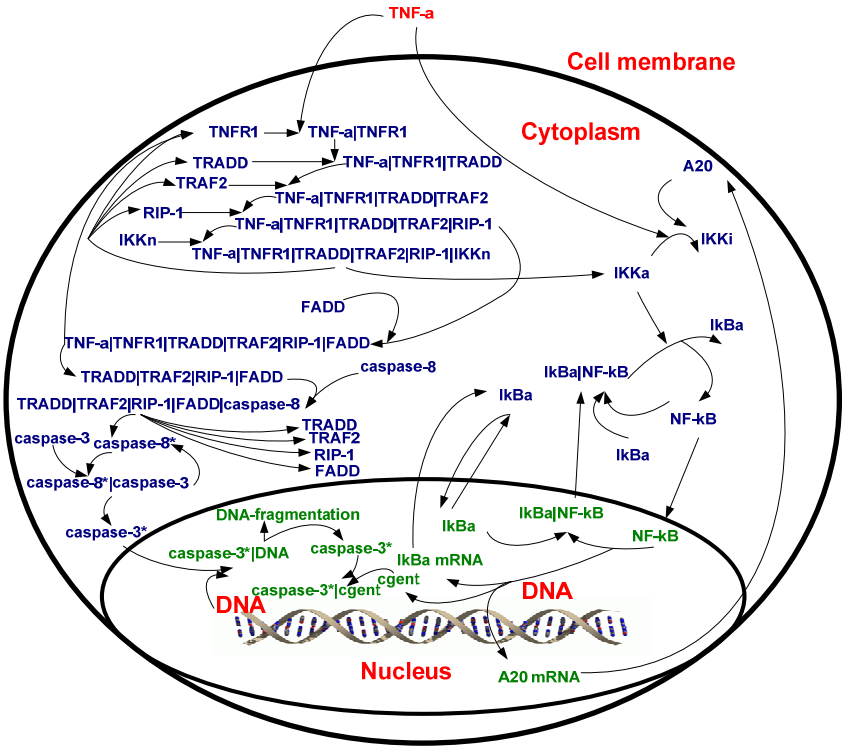


Figure 19. TNF-α signaling pathway that represents the dynamic behavior of the proteins involved in TNF-α-mediated NF-κB activation.

The integrated model, which consists of 37 differential equations and 60 parameters, can represent the dynamic behavior of the proteins involved in TNF-α-mediated NF-κB activation: TNF-α initiates the signal transduction by binding to its receptor TNFR1 and forming the complex TNF-α|TNFR1, which then recruits TRADD, TRAF2, RIP-1 to form the complex TNF-α|TNFR1|TRADD| TRAF2|RIP-1. This complex then activates two pathways: (1) it activates the apoptotic machinery by recruiting FADD; (2) it activates the NF-κB pathway by promoting the neutral form of IKK (IKKn) to the active form of IKK (IKKa). NF-κB is then released from the complex NF-κB|IκBα and translocates into the nucleus to initiate the transcription/translation process.

4.5.2 Fluorescence Intensity Profiles Obtained via Image Analysis

The activation of NF-κB in H35 reporter cells was investigated by stimulating with different TNF-α concentrations (6 ng/ml, 10 ng/ml, 13 ng/ml, and 19 ng/ml). The data was analyzed using the described image analysis procedure, resulting in the fluorescence intensity profiles shown in Fig. 20. The error bars indicated +/- one standard deviation from the mean of the measurements taken for each time point. The data in Fig. 20 shows that fluorescence decreases after ~11 hours even though the stimulus (TNF-α) is continually present, with the decrease being more pronounced at the higher concentrations. However, it is not clear if the decrease in fluorescence observed after ~11 hours of stimulation results from experimental artifacts (i.e., fluorescence photobleaching and cell death arising from cells being repeatedly exposed to UV light for imaging) or is a real biological phenomenon (i.e., consequence of change in gene expression arising due to constant stimulation with TNF-α). A better understanding of long-term activation is needed to evaluate this behavior.

4.5.3 Derivation of NF-κB Profiles from Fluorescence Intensity Profiles by Solving the Inverse Problem

The image analysis procedure returned the profile of the intensity $\hat{I}$ seen in the fluorescent microscopy images. Before $\hat{I}$ is used to derive the profile of NF-κB, the parameters C and Δ in the GFP model represented in Eq. (4.4)-(4.7), which link the concentration of activated GFP to the fluorescence intensity seen in an image, are estimated by the following procedure:

The $C_{\text{NF-κB}}$ data for cells stimulated by TNF-α=10 ng/ml in wild-type cells from the

paper by Hoffman et al., 2002, shows oscillation in NF-κB activation profile. Therefore, the second-order under-damped system shown in Fig. 18A is used to approximate NF-κB profile. The $C_{\text{NF-κB}}$ data presented in Hoffman et al., 2002, is then used to identify C, ε, ω_n and T_α in Eq. (4.14) with nonlinear least square optimization command in MATLAB, lsqnonlin. C, ε, ω_n and T_α are found to be 108 nM, 0.17, 4.49 and 0.27 respectively. Fig. 21 shows that the output of Eq. (4.14) with the estimated parameters C, ε, ω_n and T_α fits the $C_{\text{NF-κB}}$ data from Hoffman et al., 2002 well.

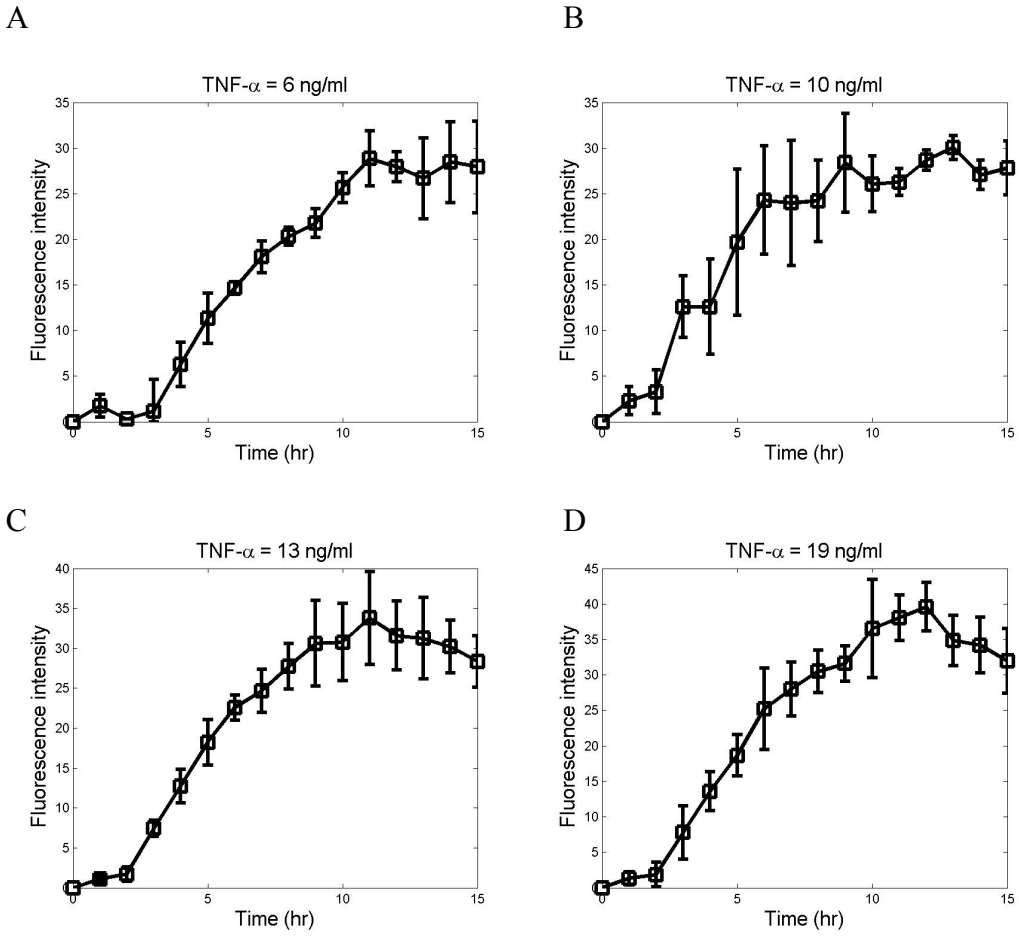


Figure 20. Fluorescence intensity profiles (A-D) obtained from the fluorescent images of the GFP reporter systems stimulated by TNF-α.

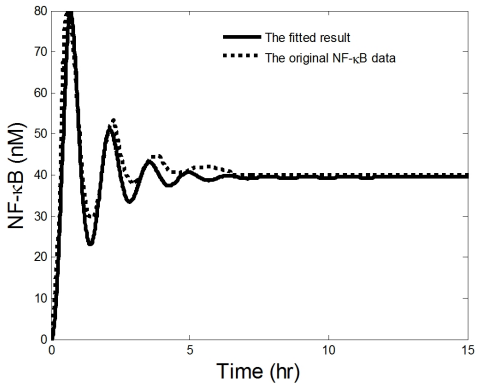


Figure 21. The output from Eq. (4.14) with the estimated parameters and the original $C_{\text{NF-κB}}$ from Hoffman et al., 2002.

The model described by Equations (4.4)-(4.7) is used with the estimated value of C, to compute the profile of fluorescence intensity. The input of this model is the concentration of NF-κB. The $C_{\text{NF-κB}}$ data for cells stimulated by TNF-α=10 ng/ml in wild-type cells from the paper by Hoffman et al., 2002, is used as an input to calculate the profile of I. As $C_{\text{NF-κB}}$ concentrations are given at discrete points, the values between two time points are estimated by linear interpolation.

The fluorescence intensity for TNF-α=10 ng/ml is computed by the described procedure from the experimental results and is shown in Fig. 22 (dash line). Δ is estimated by the ratio of the steady state value of f value computed from the model for $C_{\text{NF-κB}}$ data presented in Hoffman et al., 2002 and the steady state fluorescence intensity computed from the experimental data by the image analysis procedure. The estimated value for Δ is 2.5562×10^4.

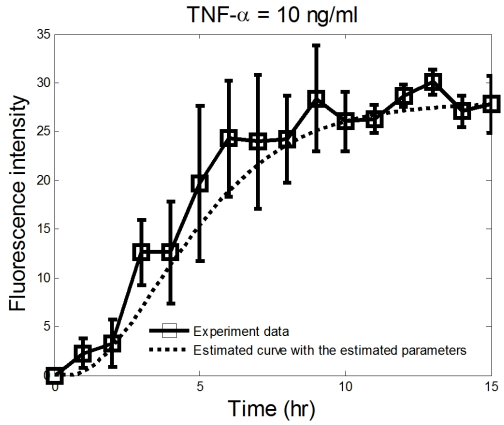


Figure 22. The experimental data and the output f/Δ from the identified GFP model for Hoffman's NF-κB data.

After C and Δ have been estimated, their values are kept constant and used to derive the profile of NF-κB from the fluorescence intensity profile $\hat{I}$ for TNF-α concentrations other than 10 ng/ml. In the following, the data for TNF-α equal to 13 ng/ml is used to illustrate the procedure of solving the inverse problem presented in Subsection 4.4.3. No assumption is made about one specific profile at this time. Instead, all three potential profiles for the input are used to determine which results in the lowest objective function value. The results are summarized in Table 7 and the corresponding graphs for the transcription factor profiles are shown in Fig. 23A.

Table 7. Values of estimated parameters for experimental data generated by stimulation with 13 ng/ml of TNF-α.

Oscillating profile (Eq. (4.14))	ε	ω_n	T_α	θ	/	Objective function
	0.25	4.5699	0.3331	0.0478	/	**4.1115**
Monotone profile (Eq. (4.17))	τ_1	τ_2	T_α	θ	/	Objective function
	0.0849	0.0551	0.3281	0.01	/	5.6976
Profile with one peak (Eq. (4.20))	p_1	z_1	T_α	τ_1	θ	Objective function
	-0.3771	-0.2684	0.4339	0.10	0.50	4.4076

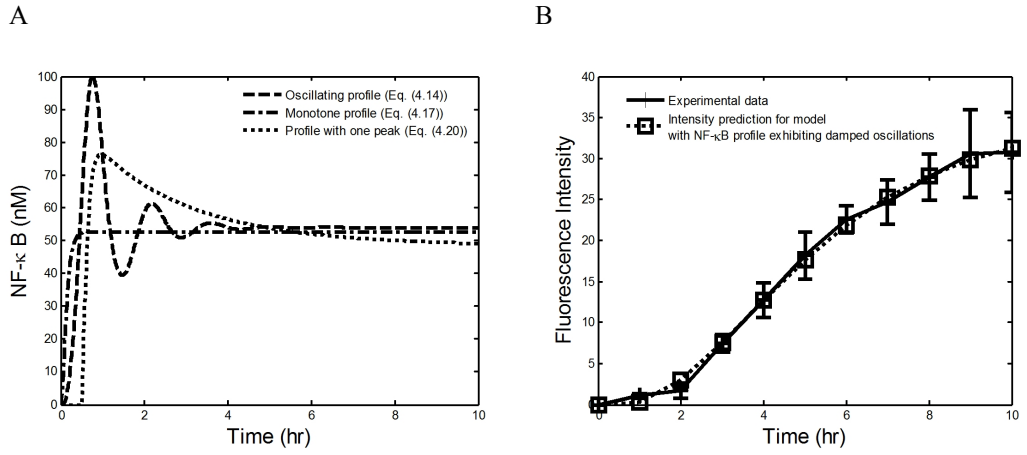


Figure 23. (A) NF-κB profiles for the experimental data generated by continuous stimulation with 13 ng/ml of TNF-α; (B) Comparison of fit between experimental and computed fluorescence intensity profile.

The results from Table 7 indicate that the NF-κB dynamics can be best estimated by damped oscillations. It can also be concluded from Fig. 23B that this profile provides an excellent fit for the experimental data. The corresponding NF-κB profile shows oscillations with an estimated period of 1.4 hour. These results are consistent with semi-quantitative Western blot data from the literature (Hoffman et al., 2002) as well as simulation results of existing models (Lipniacki et al., 2004; Rangamani and Sirovich, 2007).

As a note of caution it should be mentioned, that the objective function values shown in Table 7 are quite close to one another. This is especially true when the values for the profile exhibited damped oscillations and the one exhibiting one peak are compared. The reason for these similar objective function values is that the first peak of the oscillatory response occurs at a similar time as the peak value for the response with only one peak and that all three responses return similar long-term values for the concentration. While these results indicate that the profile exhibits damped oscillations, this can not be seen as

conclusive proof for this type of profile. Instead this technique should be used as a tool that augments the information collected from existing approaches.

Similar approach has been applied to the experiment data for the concentrations of TNF-α equal to 6 ng/ml and 19 ng/ml (refer to Fig. 20). It turns out that the NF-κB dynamics can be best estimated by damped oscillations either. The corresponding concentration profiles for NF-κB for TNF-α with concentrations of 6 ng/ml, 10 ng/ml, 13 ng/ml, and 19 ng/ml, are shown in Fig. 24. It can be seen that stimulation with higher concentrations of TNF-α results in larger long-term concentrations of NF-κB as well as in higher peak concentrations. One important aspect of this procedure is that the data obtained is quantitative (i.e., numerical values of the NF-κB profile at each time point are obtained) and not merely qualitative.

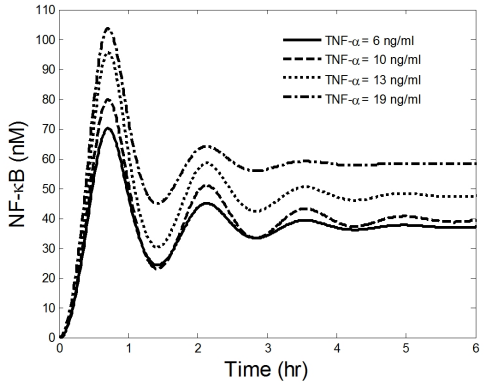


Figure 24. NF-κB profiles computed via solution of the inverse problem for TNF-α concentrations of 6 ng/ml, 10 ng/ml, 13 ng/ml, and 19 ng/ml.

4.5.4 Estimate Parameters of the Developed TNF-α Model with the Obtained NF-κB Data

These results for stimulation with 6 ng/ml, 13 ng/ml, and 19 ng/ml of TNF-α were

used to estimate parameters of the signal transduction pathway model. Since the developed model contains many more parameters than can be estimated from three time series of data, it was required to use local sensitivity analysis to determine which parameters should be re-estimated. It was determined that the parameters c_3, k_{1p}, and k_r are good candidates for estimation.

Nonlinear least square routines in MATLAB were then used to estimate these three parameters. The estimated values were found to be 0.0104, 0.0740 and 2.50, respectively. Since the data derived from the stimulation with 10 ng/ml of TNF-α was not used for estimating these parameters, this data set can be used for validating the accuracy of the updated model. Fig. 25 shows the model prediction for 10 ng/ml of TNF-α together with the experimental results derived from the described image analysis procedure. It can be concluded that the updated model predicts experimental data very well.

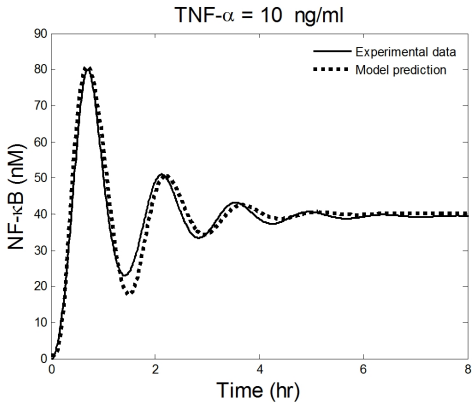


Figure 25. Comparison between NF-κB profiles computed via the presented technique for 10 ng/ml of TNF-α and updated model simulations.

4.6 Application to IL-6 Signal Transduction

Since one of the goals for this dissertation is to obtain a valid model for predicting

the dynamics of nuclear STAT3 and C/EBPβ, a model update and validation with experimental data has to be performed for the simplified IL-6 model obtained in Section 3. For this task, six sets of fluorescence intensity data were obtained for each of these two transcription factors for a step stimulation of IL6 with a concentration of 100 ng/ml, by analyzing fluorescent microscopy images of a GFP-reporter system for the corresponding transcription factor. For each set of data, there are 30 data points resulting from measurements taken every 45 minutes over a time period of 22 hours. It is possible to apply the procedure presented in Subsection 4.2-4.4 to derive the profiles of nuclear STAT3 and C/EBPβ from the fluorescent images. A simpler approach is used here as the main goal here is to re-estimate the parameters of the simplified IL-6 model instead of deriving the activation profiles of nuclear STAT3 and C/EBPβ. In this approach, the output of the simplified IL-6 model, i.e., concentrations of nuclear STAT3 or C/EBPβ, is used as the input of the GFP model shown in Eq. (4.4)-(4.7) to predict the fluorescence intensity profile that can be measured. The generated data set can then be further used to adjust parameters of the simplified IL-6 model. Fig. 26 illustrates the relationship between the simplified IL-6 model and the GFP model.

Three sets of fluorescence intensity data are used for re-estimating the parameters, while the other three data sets are used for model validation. All parameters from the simplified model, i.e., p_1-p_{19}, and S_m from the model linking the transcription factor concentration to the fluorescence intensity data, are re-estimated via the method presented in Subsection 3.2.4. The re-estimated values of the parameters of the simplified model, the equations of the model, and the initial values of the state variables

are listed in the Appendix C. The value of S_m also changed from 373 to 548 hr^{-1} due to the re-estimation.

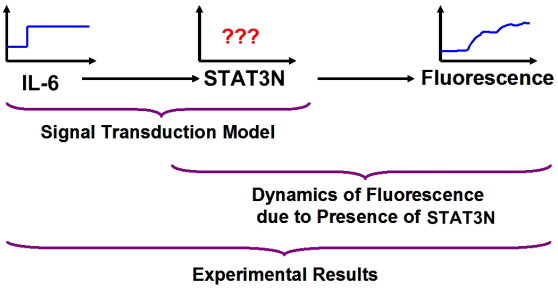


Figure 26. Relationship between input, output, and concentration of transcription factors with GFP-reporter systems.

Fig. 27 contains a comparison of one of the testing data sets with the model predictions. It can be concluded that predictions using the updated model are able to approximate the experimental data reasonably well.

A B

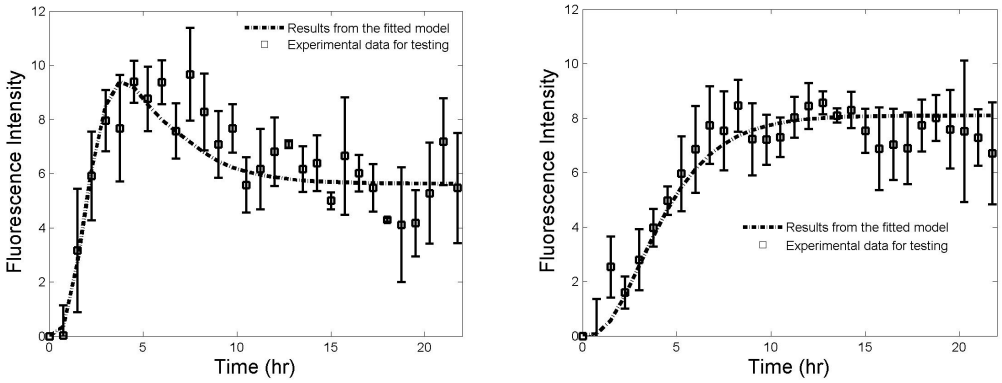


Figure 27. Comparison of fluorescence intensity profiles of the experimental data and the profiles predicted by the model. (A) Nuclear STAT3, (B) Nuclear C/EBPβ.

4.7 Summary

In this section, a methodology for quantitatively determining transcription factor profiles has been developed. This technique makes use of fluorescence microscopy images from a GFP reporter system for transcription factor activation and involves solving an inverse problem to determine the transcription factor profile from the fluorescence intensity dynamics. Data generated by this method can then be used to estimate parameters for signal transduction pathway models. This technique was applied to the activation profiles of NF-κB by TNF-α. The quantitative data of NF-κB is used to estimate the parameters in the developed model of TNF-α signaling. The presented technique was also used to derive fluorescence intensity profiles from the fluorescent images of the GFP reporter systems for nuclear STAT3 and C/EBPβ. The data was then used to re-estimated values of parameters of the simplified IL-6 model. It should be noted that the technique presented in this section can be used to determine transcription factor profiles for any system where limited qualitative knowledge about the transcription factor dynamics exists.

5. COMPUTING TRANSCRIPTION FACTOR DISTRIBUTION PROFILES FROM GREEN FLUORESCENT PROTEIN REPORTER DATA

5.1 Overview

The approach to derive quantitative data for transcription factors presented in Section 4 involves analysis of fluorescence microscopy images and solution of an inverse problem. This inverse problem deals with a model that describes changes in the fluorescence intensity due to transcription, translation, and post-translational activation of green fluorescent proteins in response to the presence of transcription factors in the nucleus. One drawback of this technique is that the approach is based upon the average fluorescence intensity over all cells and does not take into account the distribution of fluorescence intensity among a population of cells. However, inspection of fluorescence microscopy images shows that there is a clear heterogeneity in the fluorescence intensity exhibited by the cells (for an example, see Fig. 28). Additionally, it is non-trivial to predict how the fluorescence intensity will evolve over time as cells with different intensities at the beginning of an experiment may also show significantly different time-dependent profiles. One example illustrating this second point is shown in Fig. 29.

Information about phenotype heterogeneity among individual cells, i.e. the fluorescence intensity distribution in this work, plays an important role for the dynamics of the underlying signal transduction pathways (Efroni et al., 2007; Smits et al., 2005). This phenotypic heterogeneity is due to the stochasticity of the gene expression but also because of stochastic variations in the concentrations of components of the signaling

network (Raser and O'Shea, 2004). These observations form the motivation behind this work, i.e., determining transcription factor distribution profiles from experimental data. Several individual tasks need to be performed in order to achieve this goal: identification of individual fluorescent cells from fluorescence microscopy images, calculation of the fluorescence intensity distribution, and computation of the transcription factor concentration distribution from the fluorescence intensity distribution. For this purpose, individual cells are in a first step identified from fluorescence microscopy images sampled at different points in time. Next, the fluorescence intensity of each cell in the images is computed individually. The distribution of the transcription factors is then computed from the fluorescence intensity distribution.

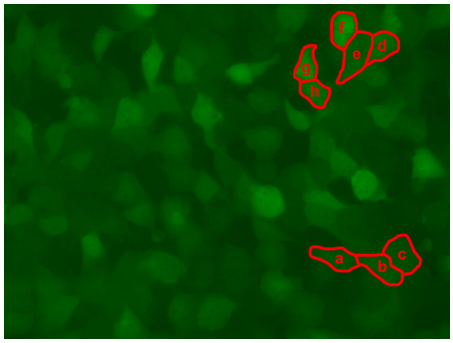


Figure 28. A fluorescence microscopy image from a NF-κB GFP reporter system stimulated by TNF-α.

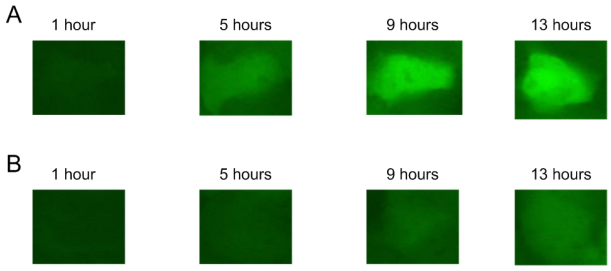


Figure 29. Time series of fluorescence microscopy images of a NF-κB GFP reporter system for two different fluorescent cells. (A) A fluorescent cell with NF-κB strongly activated, (B) A fluorescent cell with NF-κB weakly activated.

The structure of this section is as follows: Subsection 5.2 presents techniques for locating the boundary between individual fluorescent cells. Based on this, Subsection 5.3 describes an approach for quantifying the fluorescence intensity distribution of GFP reporter systems. The information of fluorescence intensity distribution is used in Subsection 5.4 to compute the distribution of transcription factor profiles. The presented techniques are applied to experimental data for the TNF-α ~ NF-κB signaling pathway in Subsection 5.5.

5.2 Derivation of the Boundary-detection Algorithm

This subsection focuses on the development of algorithms to determine the outline of fluorescent cells from images where some of the cells may border each other. In order to determine cell boundaries, the difference of the fluorescence intensity between the center region and the boundary region will be used. To describe the methodology, a portion of Fig. 28, shown in Fig. 30A, is used. The image shown in Fig. 30A contains three fluorescent cells, labeled a, b, and c. The image analysis algorithm presented in Section 4 can be applied to divide the pixels of this module into several, six in this case, different intensity levels where intensity level 1 corresponds to the highest fluorescence intensity while intensity level 6 is associated with the lowest intensity. Only pixels of intensity levels 1, 2 and 4 are shown in Fig. 30 for illustration purposes. Pixels with different fluorescence intensity levels are represented by different colors. It can be seen that the pixels with intensity level 1 of cells a and b can be found closer to the center of each cell and that the boundary region of the cells are almost devoid of this intensity level. Similar observations can be made about pixels of intensity level 2. However, some pixels with

intensity level 4 of both cells can be found next to one another. This distribution of fluorescence intensities can be used to define a center region of a cell by pixels with a high intensity, whereas the boundary region of a cell is given by pixels of lower intensity. The border between two cells will be at the boundary region of the cells. In this example, the center regions of Cell a and Cell b are separated by the pixels associated with intensity level 4. Although the areas of pixels at intensity level 4 of Cell a and b are connected, the areas of pixels of intensity level 2 of these two cells are distinct, that is, the border between Cell a and Cell b is located between the two areas of pixels of intensity level 2. Therefore, it is possible to develop an algorithm to determine the boundary between two connected fluorescent cells using the areas of pixels at different intensity levels given by the image analysis algorithm shown in Section 4.

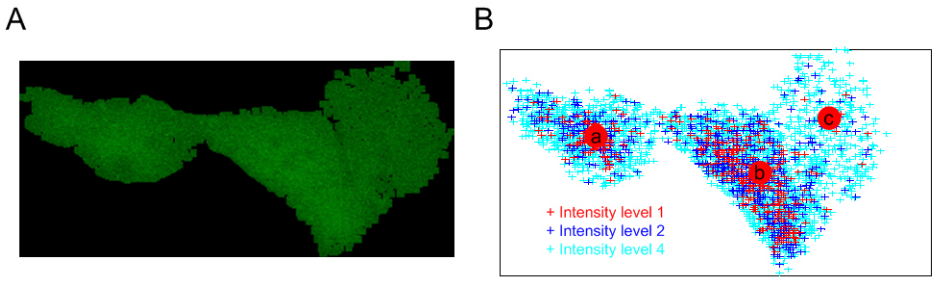


Figure 30. Example of the fluorescence intensity distribution among three adjacent cells. (A) GFP image, (B) Position distribution of fluorescence intensity. Pixels in different color represent different fluorescence intensity levels.

Three different cases need to be taken into account for computing the boundary between two fluorescent cells. These three cases depend upon the magnitude and location of the center regions of the fluorescent cells and are presented in the order of decreasing occurrence, based upon our observations: (1) two cells with center regions of

similar magnitude; (2) two cells with significant differences in the magnitude of their center regions; (3) two cells with center regions of similar magnitude but where the regions are separated by a large area of lower fluorescence intensity. The criteria that distinguish the cases from one another are also described in the following subsections.

5.2.1 Case 1: Two Cells with Center Regions of Similar Magnitude

The first scenario is given by the case where two adjacent fluorescent cells have a similar overall brightness as the number of the pixels of the first few intensity levels is similar in these two cells. For an example, refer to Cell g and Cell h in Fig. 28. For illustration purposes, the center regions of the fluorescent cells are represented as green ellipses in Fig. 31, while pixels at the lower intensity levels are marked in dark green color. The boundary of the two cells is located in the region between the outlines of the areas of Cell_1 and Cell_2.

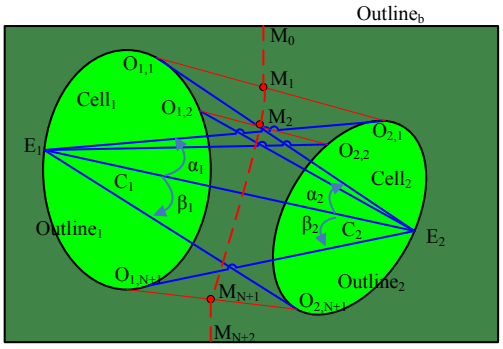


Figure 31. Illustration of the procedure for computing cell boundaries for Case 1 in which two cells have center regions of similar magnitude.

It is assumed that the pixels of the image have already been divided into n_L intensity levels, e.g., by using the image analysis algorithm presented in Section 4:

$$P_{image} = \sum_{j=1}^{n_L} P_{image,j} \tag{5.1}$$

where P_{image} is a binary image where the chosen pixels have a value of '1', and $P_{image,j}$ corresponds to the pixels having the j_{th} intensity level.

The procedure for identifying the boundary between Cell_1 and Cell_2 consists of the following steps:

Step 1: The pixels of the first n_j intensity levels, starting from 1, are labeled into separate groups (two groups for this case):

$$x_{n_j,k} = \text{bwlabel}(\sum_{j=1}^{n_j} P_{image,j}), k = 1,2 \tag{5.2}$$

where $x_{n_j,k}$ refers to the set of positions of all pixels at the first n_j intensity levels with the k_{th} label; bwlabel is a MATLAB function used for labeling connected pixels in binary images. The centers of each labeled group of pixels are then computed by the following formula

$$C_{n_j,k} = \frac{\sum_{n=1}^{n_{n_j,k}} x_{n_j,k,n}}{n_{n_j,k}}, k = 1, 2 \tag{5.3}$$

where $x_{n_j,k,n}$ represents an element in the set of $x_{n_j,k}$, and $n_{n_j,k}$ is the total number of pixels in the k_{th} labeled group of the first n_j intensity levels.

Step 2: Increase n_j until $C_{n_j-1,k}$, k =1, 2, have the same label, that is, the two separate cells belong to the same group at intensity level n_j but not at any higher intensity level. The outlines of the two ellipses in Fig. 31, can be computed by

$$Outline_k = \text{bwboundaries}(x_{n_j-1,k}), k = 1, 2 \tag{5.4}$$

The region within $Outline_k$, i.e. $x_{n_j-1,k}$, is defined as the center region of Cell_k. The outline of the pixels connecting these two cells, i.e., the outline of the dark green region in Fig. 31, is computed based on the pixels of the first n_j intensity levels

$$Outline_b = \text{bwboundaries}(x_{n_j}) \tag{5.5}$$

The areas of the center regions of the two cells, represented by $Area_1$ and $Area_2$, respectively, are defined by the number of pixels in $x_{n_j-1,k}$, k =1, 2. $C_{n_j-1,k}$, k =1, 2, are marked as C_1 and C_2 in Fig. 31.

Step 3: A straight line, drawn through the two centers C_1 and C_2, intersects the outlines of Cell_1 and Cell_2. The intersection points at the outline of Cell_1 are defined as follows:

$$\text{inter\_outline}_1 =: \{x \mid \left| \frac{(x-\text{C}_1)^{\text{T}}(x-\text{C}_2)}{\|x-\text{C}_1\|\|x-\text{C}_2\|} - 1 \right| < 0.01, x \in Outline_1\} \tag{5.6}$$

Among these intersection points, the one farthest away from C_2 is defined as E_1. E_2 is computed on the outline of Cell_2 in a similar way.

Step 4: The angles α_1 and β_1 are defined as the angles between the line E_1 - E_2 and the two tangents from E_1 to the outline of Cell_2. They are computed based upon the cross product between the vector ($\text{E}_2 - \text{E}_1$) and the one representing the line linking E_1 to any pixel on $Outline_2$.

$$\begin{aligned} &\alpha_1 = \max(\alpha), \\ &\beta_1 = \min(\alpha), \\ \alpha \coloneqq \{\arcsin(&\frac{\mathbf{V}_1 \times \mathbf{V}_2 \cdot \hat{\mathbf{n}}}{\|\mathbf{V}_1\|\|\mathbf{V}_2\|}), \mathbf{V}_1 \coloneqq \mathrm{E}_2 - \mathrm{E}_1, \mathbf{V}_2 \coloneqq x - \mathrm{E}_1, \forall x \in Outline_2\} \end{aligned} \tag{5.7}$$

where $\hat{\mathbf{n}}$ is a unit vector perpendicular to the plane spanned by $\mathbf{V}_1$ and $\mathbf{V}_2$ and in a direction given by the right-hand rule for $\mathbf{V}_1$ and $\mathbf{V}_2$. Similarly, α_2 and β_2 can be computed via the following equations:

$$\begin{aligned} &\alpha_2 = \min(\alpha), \\ &\beta_2 = \max(\alpha), \\ \alpha \coloneqq \{\arcsin(&\frac{\mathbf{V}_1 \times \mathbf{V}_2 \cdot \hat{\mathbf{n}}}{\|\mathbf{V}_1\|\|\mathbf{V}_2\|}), \mathbf{V}_1 \coloneqq \mathrm{E}_1 - \mathrm{E}_2, \mathbf{V}_2 \coloneqq x - \mathrm{E}_2, \forall x \in Outline_1\} \end{aligned} \tag{5.8}$$

An illustration of these angles is given in Fig. 31.

Step 5: The total observation angle from E_1 to $Outline_2$, which is defined as the sum of the absolute values of α_1 and β_1, is divided into N equal sections. A line starting from E_1 with a radius defined by R_1 and an angle of φ_i with the line E_1-E_2 intersects the outline of Cell_2 at point $O_{2,i}$. The equations for computing $O_{2,i}$ are as follows:

$$\begin{aligned} &Ex_{1,i} = E_1 + R_1 \begin{bmatrix} \cos(\theta + \varphi_i) \\ \sin(\theta + \varphi_i) \end{bmatrix}, \\ &R_1 = \max(\|x - E_1\|), \forall x \in Outline_2 \\ &\theta = \arctan\frac{\mathbf{V}_1(2)}{\mathbf{V}_1(1)}, \mathbf{V}_1 = E_2 - E_1 \\ &\varphi_i = \alpha_1 - (i-1)\frac{\alpha_1 - \beta_1}{N}, i = 1,2,\cdots,N+1 \end{aligned} \tag{5.9}$$

$$\begin{aligned} &O_{2,i} = x\_Ex_{1,i}\_E_1 \mid \|x\_Ex_{1,i}\_E_1 - E_1\| = \min(\|x\_Ex_{1,i}\_E_1 - E_1\|) \\ &x\_Ex_{1,i}\_E_1 =: \{x \mid \left| \frac{(x - Ex_{1,i})^{\mathrm{T}}(x - E_1)}{\|x - Ex_{1,i}\|\|x - E_1\|} + 1 \right| < 0.01, x \in Outline_2\} \end{aligned} \tag{5.10}$$

The intersection points $O_{1,i}$ can be obtained for $Outline_1$ in a similar way by using a radius R_2 and an angle φ_i with the line E_1-E_2. The corresponding equations for determining $O_{1,i}$ are:

$$\begin{aligned} Ex_{2,i} &= E_2 - R_2\begin{bmatrix}\cos(\theta+\varphi_i)\\ \sin(\theta+\varphi_i)\end{bmatrix}, \\ R_2 &= \max(\|x - E_2\|), \forall x \in Outline_1 \\ \theta &= \arctan\frac{\mathbf{V}_2(2)}{\mathbf{V}_2(1)}, \mathbf{V}_2 = E_1 - E_2 \\ \varphi_i &= \alpha_2 + (i-1)\frac{\beta_2 - \alpha_2}{N}, i = 1,2,\cdots,N+1 \end{aligned} \tag{5.11}$$

$$\begin{aligned} O_{1,i} &= x\_Ex_{2,i}\_E_2 \,|\, \|x\_Ex_{2,i}\_E_2 - E_2\| = \min(\|x\_Ex_{2,i}\_E_2 - E_2\|) \\ x\_Ex_{2,i}\_E_2 &=: \{x \,|\, \left|\frac{(x - Ex_{2,i})^{\mathrm{T}}(x - E_2)}{\|x - Ex_{2,i}\|\|x - E_2\|} + 1\right| < 0.01, x \in Outline_1\} \end{aligned} \tag{5.12}$$

Step 6: The middle of the line $O_{1,i}$- $O_{2,i}$, i =1, 2, … , N+1, is defined as M_i.

$$M_i = \frac{O_{1,i} + O_{2,i}}{2} \tag{5.13}$$

The boundary of the two cells is determined by the line passing through all M_i where i =1, 2, … , N+1.

Step 7: The line from M_1 to M_{N+1} is extended to the outline of the region where $Cell_1$ and $Cell_2$ are connected, i.e. the outline of pixels for the next lower intensity level than that of the center region, which is marked as $Outline_b$ in Fig. 31. The points of $Outline_b$ closest to the curve passing through M_1 and M_{N+1} are chosen, and called M_0 and M_{N+2}:

$$M_0 = x \,|\, \|x - M_1\| = \min(\|x - M_1\|) \forall x \in Outline_b \tag{5.14}$$

$$M_{N+2} = x \mid \|x - M_{N+1}\| = \min(\|x - M_{N+1}\|) \forall x \in Outline_b \tag{5.15}$$

Step 8: The line from M_0 to M_{N+2} defines the boundary of the two cells.

While this first case assumed that the two fluorescent cells have a similar brightness, this is not a limiting assumption as cells with significant differences in their fluorescence intensity can be easily determined with the existing algorithm presented in Section 4.

5.2.2 Case 2: Two Cells with Significant Differences in the Magnitude of Their Center Regions

This second category is characterized by a bright cell adjacent to a dark cell where the cell with the higher intensity level has significantly more bright pixels than the other one. For an example of this case, refer to Cell e and Cell f in Fig. 28. Parts of the shared boundary region in Fig. 32 belong to $Cell_2$ because the pixels from the center region of the darker cell may have a similar intensity level as the ones found in the boundary region of the brighter cell. In this case, parts of the outline of $Cell_1$ are used as the boundary between these two cells to avoid that parts of $Cell_2$ are incorrectly attributed to $Cell_1$.

The procedure for identifying the boundary of $Cell_1$ and $Cell_2$ for Case 2 consists of the following steps:

Step 1: Step 1 ~ 3 as described for Case 1 are performed to obtain the outlines and center regions of the two fluorescent cells.

Step 2: If the ratio of the area of the center region of $Cell_1$ to that of $Cell_2$ is larger than a threshold value, i.e. r_1 shown in Eq. (5.16), then the boundary between the two cells is considered to be of Case 2 type. In this work r_1 is set to a value of 10.

$$\frac{Area_1}{Area_2} > r_1 \tag{5.16}$$

Step 3: Steps 4 ~ 5 as described for Case 1 are performed. However, only the intersection points that are on the outline of $Cell_1$, i.e., $O_{1,i}$, i =1, 2,… ,N+1, are computed.

Step 4: The outline section from $O_{1,1}$ to $O_{1,N+1}$ of $Cell_1$, in a clock-wise direction, is used as the boundary between $Cell_1$ and $Cell_2$. The outline section from $O_{1,1}$ to $O_{1,N+1}$ is then extend to $O_{1,0}$ and $O_{1,N+2}$ in a similar manner as described in Step 7 of Case 1.

Step 5: The curve from $O_{1,0}$ to $O_{1,N+2}$ is considered to be the boundary of the two cells.

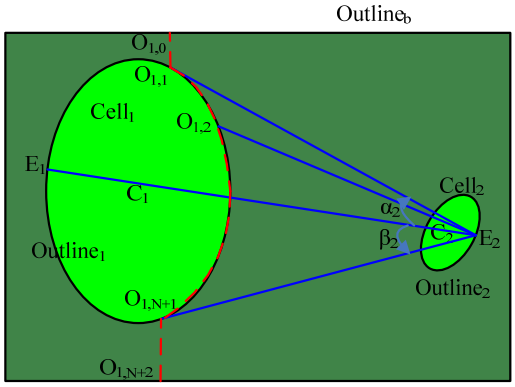


Figure 32. Illustration of the procedure for computing cell boundaries for Case 2 in which two cells have significant differences in the magnitude of their center regions.

5.2.3 Case 3: Two Cells with Center Regions of Similar Magnitude but Where the Regions Are Separated by a Large Area of Lower Fluorescence Intensity

This case mainly occurs if a cell with very low fluorescence intensity is located between two cells with significantly higher brightness. One example of this type is given

by Cells d and f in Fig. 28 as cell e is located between Cells d and f, but Cell e has lower fluorescence intensity than either d or f. If this situation occurs then the following procedure is applied to obtain the boundaries of these cells:

Step 1: Step 1 ~ 3 as described for Case 1 are performed to obtain the outlines and center regions of the two fluorescent cells.

Step 2: The shortest distance between the outlines of Cell_1 and Cell_2 is computed, i.e. *D_o* in Fig. 33, using the following equation:

$$D\_o = \min(\min(\|x_1 - x_2\|), \forall x_2 \in \text{Outline}_2), \forall x_1 \in \text{Outline}_1 \tag{5.17}$$

Step 3: If the ratio of *D_o* divided by the distance of the centers of the two cells is larger than a pre-specified threshold value, i.e. r_2 shown in Eq. (5.18), then Case 3 is considered for these two cells. In this work, r_2 is set to a value of 0.3.

$$\frac{D\_o}{|C_1 - C_2|} > r_2 \tag{5.18}$$

Step 4: Steps 4 ~ 5 as described for Case 1 are performed to obtain the intersection points located on the outlines of the center regions of the two cells, i.e., $O_{1,i}$ and $O_{2,i}$, i =1, 2,… , N+1,.

Step 5: Instead of using the middle point of the line $O_{1,i}$-$O_{2,i}$, i =1, 2,… , N+1, as the boundary of these two cells, the curve from $O_{1,1}$ to $O_{1,N+1}$ in a clock-wise direction for Cell_1 and the one from $O_{2,1}$ to $O_{2,N+1}$ in a counter clock-wise direction for Cell_2 are taken as the boundaries. The rationale for this step is that Cell_3, which is located in between Cell_1 and Cell_2, would be cut into two pieces if the same approach as the one described in Case 1 would be applied here.

Step 6: The boundaries obtained in Step 5 are extended to *Outline*$_b$ as described in Step 7 of Case 1.

Step 7: The two curves, one including the points from $O_{1,0}$ to $O_{1,n+1}$ and the other including the points from is $O_{2,0}$ to $O_{2,n+1}$, are considered to be the boundaries of the two cells.

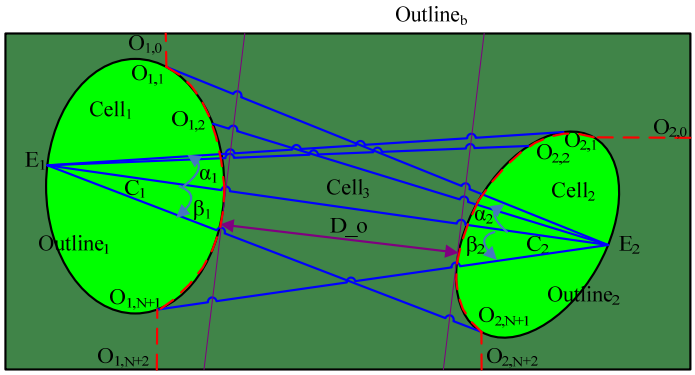


Figure 33. Illustration of the procedure for computing cell boundaries for Case 3 in which two cells have center regions of similar magnitude but are separated by a large area of lower fluorescence intensity.

5.3 Image Analysis Algorithm for Identifying Individual Cells from Fluorescent Images

While Subsection 5.2 introduced procedures that can used to determine the boundaries of individual fluorescent cells if the center regions of two cells have been clearly identified, it is necessary to first extract this type of information from the fluorescent images. This task is exacerbated because several fluorescent cells are often found directly adjacent to each other (which will be called fluorescent cell modules in the following text). Furthermore, some pixels with a high fluorescence intensity may represent noise.

To address these points, this section presents a mathematical morphology approach to remove isolated noisy pixels and obtain an initial outline of fluorescent cell modules

via an opening operation. Based upon the boundary detection algorithm presented in Subsection 5.2, a procedure is presented to iteratively divide each fluorescent cell module to obtain the outlines of individual fluorescent cells within the modules. This algorithm consists of the following steps:

Step 1: Analyze the fluorescence microscopy image with the algorithm based upon PCA and K-means-clustering presented by Section 4 to obtain the results shown in Eq. (5.1). The pixels for the first N_f intensity levels are considered to represent the fluorescent cell region P_{image_f}.

$$P_{image\_f} = \sum_{j=1}^{N_f} P_{image,j} \tag{5.19}$$

Step 2: Apply a mathematical morphology operation to remove the isolated noisy pixels in P_{image_f} and then separate the image into unconnected fluorescent cell modules via the following steps:

Step 2.1: Overlay a 3×3 square structuring element on P_{image_f}, and define pixels that have less than three '1' pixels in the structuring element as isolated pixels. Remove these pixels by setting their values to 0.

Step 2.2: Apply an opening operation with a disk structuring element to the image generated by Step 2.1. The image generated by Steps 2.1~2.2 is defined as $P_{image\_f\_denoise}$.

Step 2.3: The pixels of a fluorescent cell region are given the same label using the MATLAB command 'bwlabel'. $P_{image\_f\_denoise}$ is divided into several labeled fluorescent cell modules.

$$P_{image\_f\_denoise} = \sum_{i=1}^{num\_labeled} P_{image\_f\_denoise,i} \tag{5.20}$$

Step 3: Iteratively apply the boundary identification algorithm presented in Subsection 3.1 to obtain the individual cells in each labeled fluorescent cell region $P_{image\_f\_denoise,i}$, via the following procedures:

Step 3.1: Apply the image analysis algorithm based upon PCA and K-means clustering to the pixels of each region, $P_{image\_f\_denoise,i}$.

$$P_{image\_f\_denoise,i} = \sum_{j=1}^{n_L} P_{image\_f\_denoise,i,j} \tag{5.21}$$

where $P_{image\_f\_denoise,i,j}$ represents the pixels of the j_{th} intensity level of $P_{image\_f\_denoise,i}$.

Step 3.2: Obtain the fluorescent cell boundaries for each $P_{image\_f\_denoise,i,}$ based upon the boundary detection algorithm presented in Subsection 5.2 via steps 3.2.1~3.2.8:

Step 3.2.1: Identify the pixels having one of the first m_c fluorescence intensity levels:

$$S_{m_c} = \sum_{j=1}^{m_c} P_{image\_f\_denoise,i,j} \tag{5.22}$$

Step 3.2.2: Divide S_{m_c} into unconnected fluorescent cell modules $S_{m_c,k}$, k = 1, 2, …, n_{m_c} where n_{m_c} represents the number of fluorescent cell modules in S_{m_c}, and obtain the morphological properties such as the center $C_{S_{m_c,k}}$ and the outline $O_{S_{m_c,k}}$ of each labeled region. The MATLAB functions ‘bwlabel’ and ‘regionprops’ are used for these operations.

$$S_{m_c} = \sum_{k=1}^{n_{m_c}} S_{m_c,k} \tag{5.23}$$

Step 3.2.3: Determine whether a labeled region for the first m_c-1 intensity levels, i.e. $S_{m_c\text{-}1,k}$ (k begins with 1 and stop at $n_{m_c\text{-}1}$-1), is connected to other labeled regions $S_{m_c\text{-}1,l}$, l = k+1, k+2, ..., $n_{m_c\text{-}1}$ at the m_c intensity level: if the centers of $S_{m_c\text{-}1,k}$ and $S_{m_c\text{-}1,l}$, i.e. $C_{S_{m_c\text{-}1,k}}$ and $C_{S_{m_c\text{-}1,l}}$, belong to the same labeled region $S_{m_c,q}$ (q=1, 2,... , n_{m_c}), $S_{m_c\text{-}1,k}$ is connected to $S_{m_c\text{-}1,l}$ in S_{m_c}. The set of labeled regions that $S_{m_c\text{-}1,k}$ is adjacent to in S_{m_c} can be represented by a set $U_{S_{m_c\text{-}1,k}}$ where

$$U_{S_{m_c\text{-}1,k}} =: \{S_{m_c\text{-}1,l} \mid C_{S_{m_c\text{-}1,k}} \in S_{m_c,q}, \text{and } C_{S_{m_c\text{-}1,l}} \in S_{m_c,q}, q \in \{1,2,\cdots, \ n_{m_c}\}\} \tag{5.24}$$

Step 3.2.4: Determine whether there is a boundary between $S_{m_c\text{-}1,k}$ and every element of $U_{S_{m_c\text{-}1,k}}$. This task can be performed as follows: (1) The elements in $U_{S_{m_c\text{-}1,k}}$, represented as $U^{(q)}_{S_{m_c\text{-}1,k}}$, q =1, 2, ⋯, n_U, are indexed according to the distance between the center of $S_{m_c\text{-}1,k}$, i.e. $C_{S_{m_c\text{-}1,k}}$, and the centers of the elements of $U_{S_{m_c\text{-}1,k}}$ in an increasing order, that is, $C_{U^{(1)}_{S_{m_c\text{-}1,k}}}$ represents the center of the element whose center is closest to $C_{S_{m_c\text{-}1,k}}$; (2) The two tangents from the center of $S_{m_c\text{-}1,k}$ to the outline of each element in $U_{S_{m_c\text{-}1,k}}$, which is given by Steps 3.2.1 and 3.2.2, are marked as $\mathrm{T}_{S_{m_c\text{-}1,k},U^{(q)}_{S_{m_c\text{-}1,k}},\alpha}$ and $\mathrm{T}_{S_{m_c\text{-}1,k},U^{(q)}_{S_{m_c\text{-}1,k}},\beta}$, q = 1, 2, …, n_U, respectively. These are computed using a similar approach, given by Eq. (5.7); (3) $U^{(1)}_{S_{m_c\text{-}1,k}}$ is kept in $U_{S_{m_c\text{-}1,k}}$ while any other element is removed from $U_{S_{m_c\text{-}1,k}}$ if the line from $C_{S_{m_c\text{-}1,k}}$ to $C_{U^{(1)}_{S_{m_c\text{-}1,k}}}$ falls between

the two tangents from $C_{S_{m_c\text{-}1,k}}$ to its outline or the line linking $C_{S_{m_c\text{-}1,k}}$ to its center falls between $\mathrm{T}_{S_{m_c\text{-}1,k},U^{(1)}_{S_{m_c\text{-}1,k}},\alpha}$ and $\mathrm{T}_{S_{m_c\text{-}1,k},U^{(1)}_{S_{m_c\text{-}1,k}},\beta}$. Mathematically, these two conditions for removing the q_{th} element from $U_{S_{m_c\text{-}1,k}}$ are represented by Eq. (5.25) and (5.26).

$$\mathrm{sgn}(\mathbf{V}_1 \times \mathrm{T}_{S_{m_c\text{-}1,k},U^{(q)}_{S_{m_c\text{-}1,k}},\alpha} \cdot \hat{\mathbf{n}}_1)\mathrm{sgn}(\mathbf{V}_1 \times \mathrm{T}_{S_{m_c\text{-}1,k},U^{(q)}_{S_{m_c\text{-}1,k}},\beta} \cdot \hat{\mathbf{n}}_2) < 0 \tag{5.25}$$
$$\mathbf{V}_1 := C_{U^{(1)}_{S_{m_c\text{-}1,k}}} - C_{S_{m_c\text{-}1,k}}$$

$$\mathrm{sgn}(\mathbf{V}_2 \times \mathrm{T}_{S_{m_c\text{-}1,k},U^{(1)}_{S_{m_c\text{-}1,k}},\alpha} \cdot \hat{\mathbf{n}}_3)\mathrm{sgn}(\mathbf{V}_2 \times \mathrm{T}_{S_{m_c\text{-}1,k},U^{(1)}_{S_{m_c\text{-}1,k}},\beta} \cdot \hat{\mathbf{n}}_4) < 0 \tag{5.26}$$
$$\mathbf{V}_2 := C_{U^{(q)}_{S_{m_c\text{-}1,k}}} - C_{S_{m_c\text{-}1,k}},$$

where sgn is the sign function, $\hat{\mathbf{n}}_1$ is a unit vector perpendicular to the plane spanned by $\mathbf{V}_1$ and $\mathrm{T}_{S_{m_c\text{-}1,k},U^{(q)}_{S_{m_c\text{-}1,k}},\alpha}$ and in a direction given by the right-hand rule for $\mathbf{V}_1$ and $\mathrm{T}_{S_{m_c\text{-}1,k},U^{(q)}_{S_{m_c\text{-}1,k}},\alpha}$, other $\hat{\mathbf{n}}_q$, q = 2, 3, 4 are obtained in a similar way.

Step 3.2.5: Apply the boundary detection algorithm presented in Subsection 5.2 to $S_{m_c\text{-}1,k}$ and each element remaining in $U_{S_{m_c\text{-}1,k}}$, i.e., $U^{(q)}_{S_{m_c\text{-}1,k}}$. Specifically, the pixels in $S_{m_c\text{-}1,k}$ and $U^{(q)}_{S_{m_c\text{-}1,k}}$ of $U_{S_{m_c\text{-}1,k}}$ make up the center regions for these two fluorescent cells or fluorescent cell modules, while the outlines of $S_{m_c\text{-}1,k}$ and $U^{(q)}_{S_{m_c\text{-}1,k}}$ can be used for the outlines given by Eq. (5.4). The boundary between $S_{m_c\text{-}1,k}$ and $U^{(q)}_{S_{m_c\text{-}1,k}}$ is then determined via the approach shown in Subsection 5.2.

Step 3.2.6: The pixels along the boundary identified in Step 3.2.5 for S_{m_c} are removed from $P_{image\_f\_denoise,i}$.

Step 3.2.7: Increase k by one and repeatedly performs Steps 3.2.3 ~ 3.2.6 on $S_{m_c\text{-}1,k}$.

Step 3.2.8: Increase m_c by one and repeat Step 3.2.1~ 3.2.7 until the boundaries have been found for all intensity levels, i.e., $U_{S_{m_c\text{-}1,k}}$ is an empty set for all labeled regions in $S_{m_c\text{-}1}$.

Step 3.3: Obtain the indices of fluorescent cells in $P_{image\_f\_denoise,i,}$ with the MATLAB command 'bwlabel'.

Step 3.4: Check whether the fluorescent cell regions computed in Steps 3.2~3.3 are individual fluorescent cells or still regions representing smaller fluorescent cell modules containing several cells. The reason for performing this procedure is that some adjacent cells may have a similar fluorescence intensity level. These fluorescent cell modules can be further identified based upon the following characteristics: their sizes are usually larger than the size of a normal single cell (e.g., a size exceeding 5000 pixels) and their shapes are distinctively different from a circle, e.g., the difference of the major and minor axis length is larger than half of the equivalent diameter.

Step 3.5: For each cell region that needs to be further divided, the pixels within this region are analyzed using the same procedure that has been applied to $P_{image\_f\_denoise,i}$ in Steps 3.1~ 3.4.

Step 4: The individual cells in the entire image are given by the computed fluorescent cells for each labeled fluorescent cell module $P_{image\_f\_denoise,i}$, i = 1, 2, …, *num_labeled*.

After the individual fluorescent cells have been identified by the above described

algorithm, it is possible to compute the fluorescence intensity of individual cells and thereby the fluorescence intensity distribution. The presented algorithm is applied to a series of images taken at different points in time resulting in a data set that describes the fluorescence intensity distribution of a collection of cells over time.

5.4 Deriving the Transcription Factor Distribution from the Fluorescence Intensity Distribution

As mentioned in the overview subsection, it can be observed in experiments that there is a distribution of the fluorescence intensity among different cells of a population. This heterogeneity of the fluorescence intensity is caused in part by a distribution of the transcription factor concentrations (Smits et al., 2006). Based upon the observation that cells exhibiting higher fluorescence intensity at the beginning of an experiment will likely continue to be brighter than the ones that started out with a lower intensity, e.g., see Fig. 29, it can be concluded that the activation of a transcription factor in individual fluorescent cells can be appropriately described by the bulk average activation of the transcription factor over all fluorescent cells in the entire image (represented as $\overline{C}_{TF}$) multiplied by a stochastic coefficient. This subsection presents a procedure to derive the transcription factor concentration distribution from the fluorescence intensity distribution by solving an inverse problem for the GFP model given by Eq. (4.5) - (4.7) and (4.9). The procedure consists of the following steps:

Step 1: Compute $\overline{C}_{TF}(t)$ via the approach presented in Section 4 and then compute $\overline{u}(t)$ using Eq. (4.8);

Step 2: Estimate the probability density functions, $f_{\text{pdf\_experiment}}(t_i)$, from the

fluorescence intensity histograms of individual cells in the images measured at t_i, i=1,2,⋯, N_t.

Step 3: The input of the GFP model, u in Eq. (4.9), is approximated by the product of $\bar{u}(t)$ and a randomly generated coefficient from a log-normal distribution with the variance σ and mean μ:

$$u(t) = \bar{u}(t)\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \tag{5.27}$$

N samples of u are generated from Eq. (5.27) with parameters σ and μ using the following procedure: (1) the cumulative distribution function f_{cdf} is obtained from the probability distribution function; (2) a random number r_{uniform} is generated from a uniform distribution; (3) one sample of u is obtained by multiplying $\bar{u}$ with a coefficient a which satisfies the following relationship: $f_{\text{cdf}}(a) = r_{\text{uniform}}$; (4) the above three steps are repeated N times such that N samples of u are obtained.

Step 4: For each sample of u, the corresponding fluorescence intensity profile I is computed via the GFP model given by Eq. (4.5) - (4.7), and (4.9). N samples of I are obtained corresponding to the N samples of u.

Step 5: The probability density functions, $f_{\text{pdf\_model}}(t_i)$, for I at time t_i, i=1,2,⋯, N_t, are estimated from the corresponding histograms of I obtained in Step 4.

Step 6: The objective function used for estimation of σ and μ is given by the sum of the squared differences between $f_{\text{pdf\_experiment}}(t_i)$ and $f_{\text{pdf\_model}}(t_i)$ over all times t_i, i=1,2,⋯, N_t.

$$\min_{\sigma,\mu} \sum_{i=1}^{N_t} (f_{\text{pdf\_experiment}}(t_i) - f_{\text{pdf\_model}}(t_i))^2 \tag{5.28}$$

subject to

for one sample of u :

$$u(t) = \bar{u}(t) \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

$$I = \tilde{g}(u(t))$$

for N samples *of* u :

$$f_{\text{pdf\_model}}(t_i) = pdf(I(t_i))$$

where $\tilde{g}$ represents the model describing GFP dynamics, i.e. the model consisting of Equatios (4.5) - (4.7), and (4.9); the function $pdf(I(t_i))$ refers to the probability density function of the samples of $I(t_i)$.

Step 7: A nonlinear least squares optimization technique from MATLAB, lsqnonlin, is used to minimize the objective function shown in Eq. (5.28) to determine the values for σ and μ.

Step 8*:* N samples of u are generated using Eq. (5.27) for the estimated values of σ and μ, and the distribution of C_{TF} is obtained from the N samples of C_{TF} computed by Eq. (4.10) corresponding to the N samples of u.

This procedure can be used to compute the transcription factor concentration distribution profiles from the fluorescence intensity distribution data collected at different points in time.

5.5 Application of the Procedure to Images Generated from NF-κB GFP Reporter Systems Stimulated by TNF-α

This subsection applies the described procedure to experimental data to determine a

transcription factor concentration distribution profile. The system under investigation consists of H35 cells which have been modified to serve as NF-κB GFP reporters. In the experiments the cells are continuously stimulated with 13 ng/ml of TNF-α. Three fluorescent images are taken every hour at different positions for a period of 15 hours, resulting in three data sets.

The techniques presented in Subsection 5.2-5.4 are use to derive the time-dependent distribution of NF-κB from the data sets. Subsection 5.5.1 describes how the presented image analysis technique is applied to identify individual fluorescent cells for a series of fluorescence microscopy images and to compute the distribution of fluorescence intensity at different points in time. Based upon the obtained fluorescence intensity distribution data, Subsection 5.5.2 then uses the approach shown in Subsection 5.4 to derive the NF-κB distribution from the fluorescence intensity distribution.

5.5.1 Compute the Fluorescence Intensity Distribution from Fluorescence Microscopy Images

In this subsection, the algorithms for identifying individual cells from fluorescence microscopy images are applied to the images taken of the NF-κB GFP reporter system. Fig. 28, one typical image from the data set, is used to illustrate the application of the algorithm to experimental data. In a first step, the pixels that correspond to each individual cell are identified and then the average fluorescence intensity of each cell is computed from these pixels. The fluorescence intensity distribution is then obtained from the fluorescence intensities of all individual cells. This procedure is applied to the fluorescent images taken at different points in time in order to obtain a fluorescence

intensity distribution profile.

Seven clusters of cells are obtained from the image shown in Fig. 28 based upon their fluorescence intensity. The pixels found in the first five clusters are shown in Fig. 34A. Steps 2.1 ~ 2.3 from Subsection 5.3 are applied to Fig. 34A, and the labeled fluorescent cell regions are represented in different colors in Fig. 34B. It can be seen that several fluorescent cell modules are found in the image and that almost each of these cell modules consists of several fluorescent cells. Step 3 of the procedure described in Subsection 5.3 is illustrated by applying it to the largest fluorescent cell module from Fig. 34B, i.e. the one shown in Fig. 35A.

A B

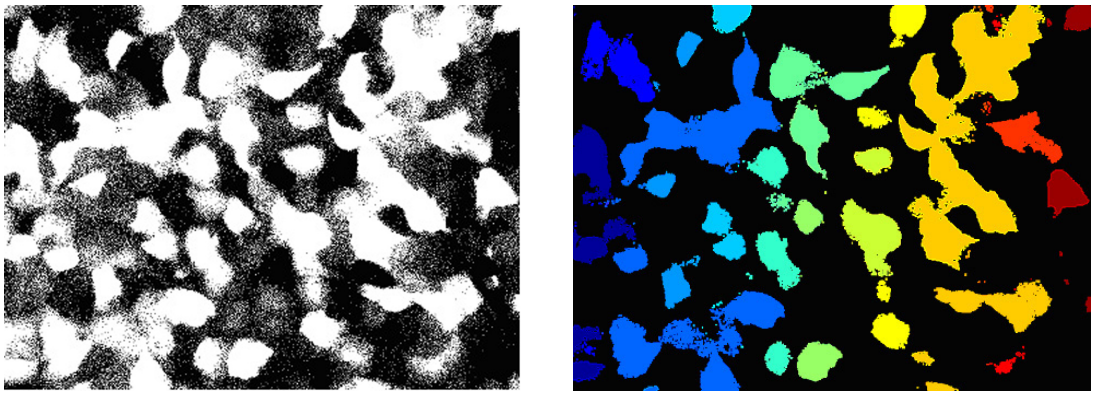


Figure 34. (A) Pixels for the first five clusters; (B) Labeled fluorescent cell modules in the de-noised image.

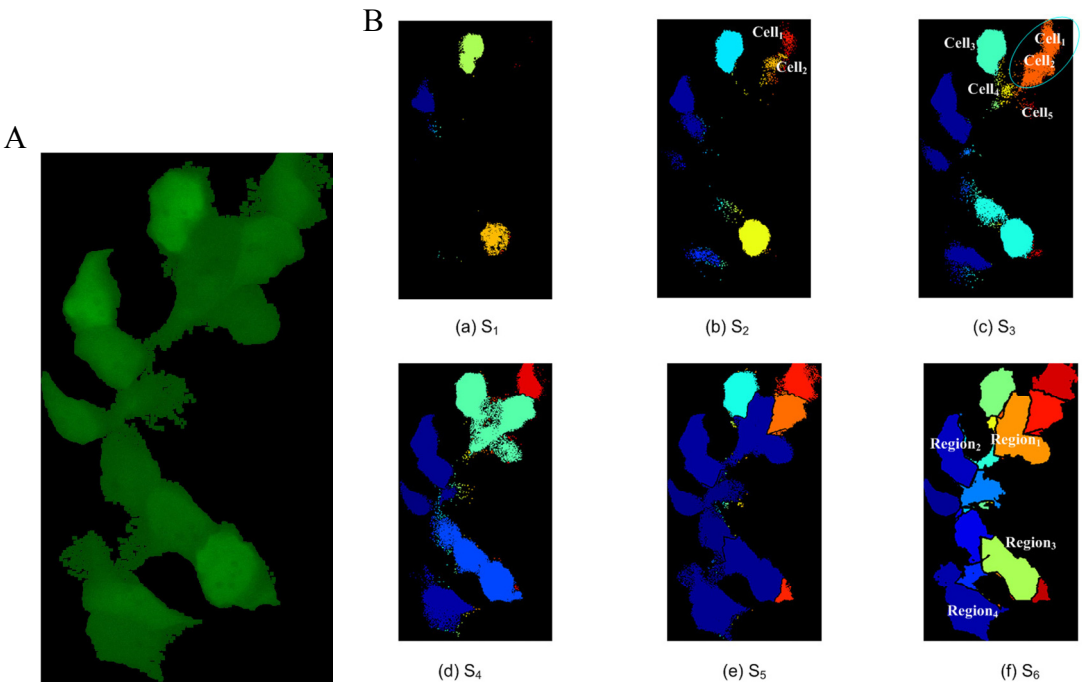


Figure 35. (A) The largest connected fluorescence region from Fig. 34B; (B) Labeled pixels correspond to S_{m_c}, m_c =1, 2, ⋯, 6; figures (a) through (f) correspond to the pixels obtained for the six different values of m_c via the algorithm described in Subsection 5.3.

The pixels of Fig. 35A are grouped into seven clusters via image analysis based upon PCA and K-means-clustering described in Step 3.1 presented in Subsection 5.3. Pixels for the first m_c intensity levels, i.e. S_{m_c}, m_c = 1, 2, …, 6, are shown in Fig. 35B. No cells are connected to other cells for m_c less than 3. Cell_1 is connected to Cell_2 for m_c equal to 3. The boundary detection algorithm for Case 1 from Subsection 5.2 is used to divide these two cells. Pixels for $\text{Cell}_2 \sim \text{Cell}_5$ are connected to each other for a value of m_c of 4. Based upon Step 3.2.5 presented in Subsection 5.3, there is a cell boundary located between the center regions of Cell_3 and Cell_4 while Cell_3 and Cell_2 should be separated by another type of boundary. The algorithm described for Case 2 from Subsection 5.2 is used to determine the boundary between Cell_3 and Cell_4 as the center region for Cell_3

shown in Fig. 35B.(c) is much larger than that of $Cell_4$. The same situation also arises for $Cell_2$ and $Cell_4$. On the other hand, the algorithm for Case 3 from Subsection 5.2 is applied to determine the boundary for $Cell_3$ and $Cell_2$ as these two cells are separated by a cell exhibiting a lower fluorescence intensity. There is no boundary between $Cell_4$ and $Cell_5$ because the center regions of these two cells are small (less than 1000 pixels) and it is possible that they belong to different parts of the same cell. However, if they were not part of the same cell, then the pixels could be assigned to different cells using the procedure described in Steps 3.4 and 3.5 presented in Subsection 5.3.

Four regions need to be further divided in Fig. 35B.(f) based upon the criteria stated in Step 3.4 presented in Subsection 5.3. The procedure described in Step 3.5 is used to identify the fluorescent cells in these specific regions and the results are shown in Fig. 36. The same procedures that have been applied to the module shown in Fig. 35A are applied to all other labeled region from Fig. 34B, i.e. for $P_{image\_f\_denoise,i,}$ with i >1. The individual cells identified from Fig. 28 are shown in Fig. 37.

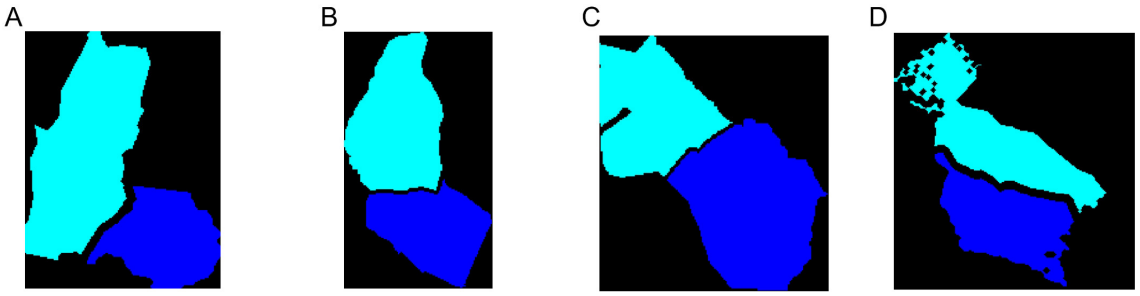


Figure 36. Cell separation for Region 1~ 4 in Fig. 35B.(f). (A) Region 1, (B) Region 2, (C) Region 3, (D) Region 4. These regions are characterized by a shape that is quite different from a circle.

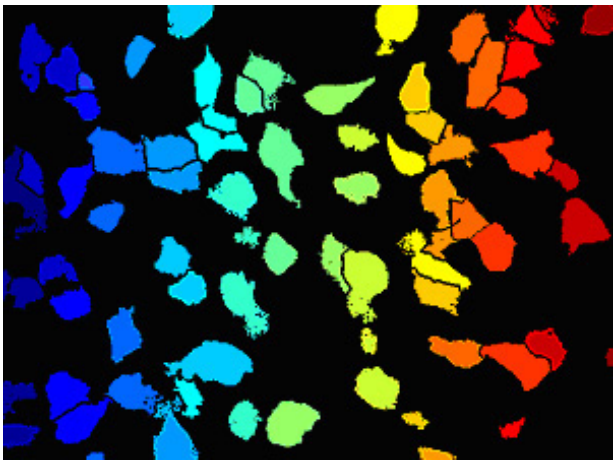


Figure 37. Individual cells identified from Fig. 28 with the techniques presented. Regions in different color represent different individual cells.

After the individual fluorescent cells have been identified from the fluorescent images using the presented procedure, a histogram of the fluorescence intensities of the individual cells is computed. Fig. 38 shows the fluorescence intensity distributions from one set of experimental data for measurements taken after 1, 5, 9, and 13 hours. While more data is available, the results are not shown here due to space constraints.

5.5.2 Compute NF-κB Distribution Profiles from Fluorescence Intensity Profiles

The presented algorithm is applied to obtain the distribution of NF-κB from the three sets of experimental data. The fluorescence intensity distribution data shown in Fig. 38 are used to estimate the parameters σ and μ. The estimated value of σ is 0.29 while the one of μ is -0.36. The distribution of NF-κB at the 1^{st}, 2^{nd}, 3^{rd}, 5^{th}, 9^{th} and 13^{th} hour is shown in Fig. 39. It can also be seen from Fig. 39 that: (1) the value of NF-κB that is most commonly found in the cells after approximately 1 hour is larger than the ones for later time periods; (2) the value of NF-κB that is most commonly found in the cells after

approximately 3 hours is smaller than the one recorded at later time points; (3) the distribution of NF-κB after approximately 5 hours does not change significantly. These results are consistent with the available data for NF-κB in the literature (Lee et al., 2000; Hoffmann et al., 2002), which show that the profile of NF-κB exhibits a peak within the first hour and then shows damped oscillatory behavior until it reaches a steady state after approximately 5 hours.

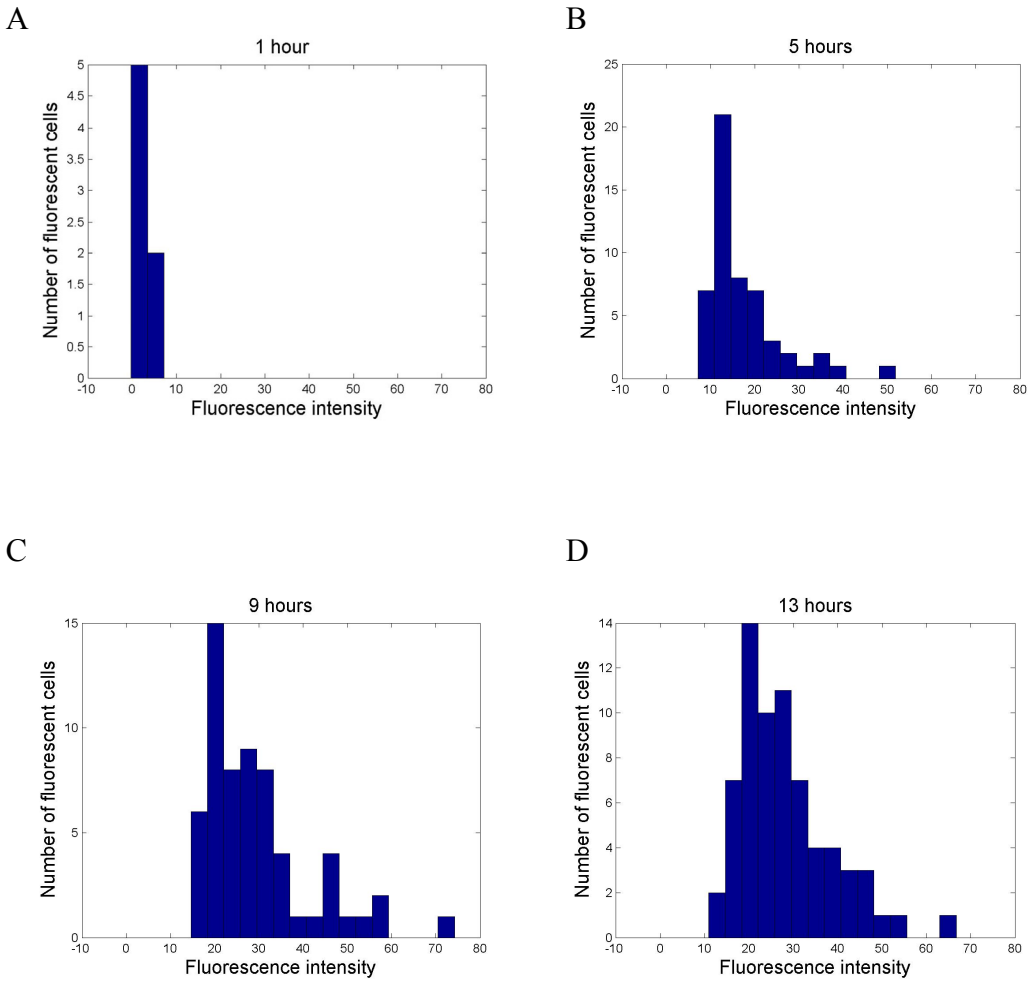


Figure 38. Fluorescence intensity distribution of cells for one set of experimental data for constant stimulation with 13 ng/ml of TNF-α.

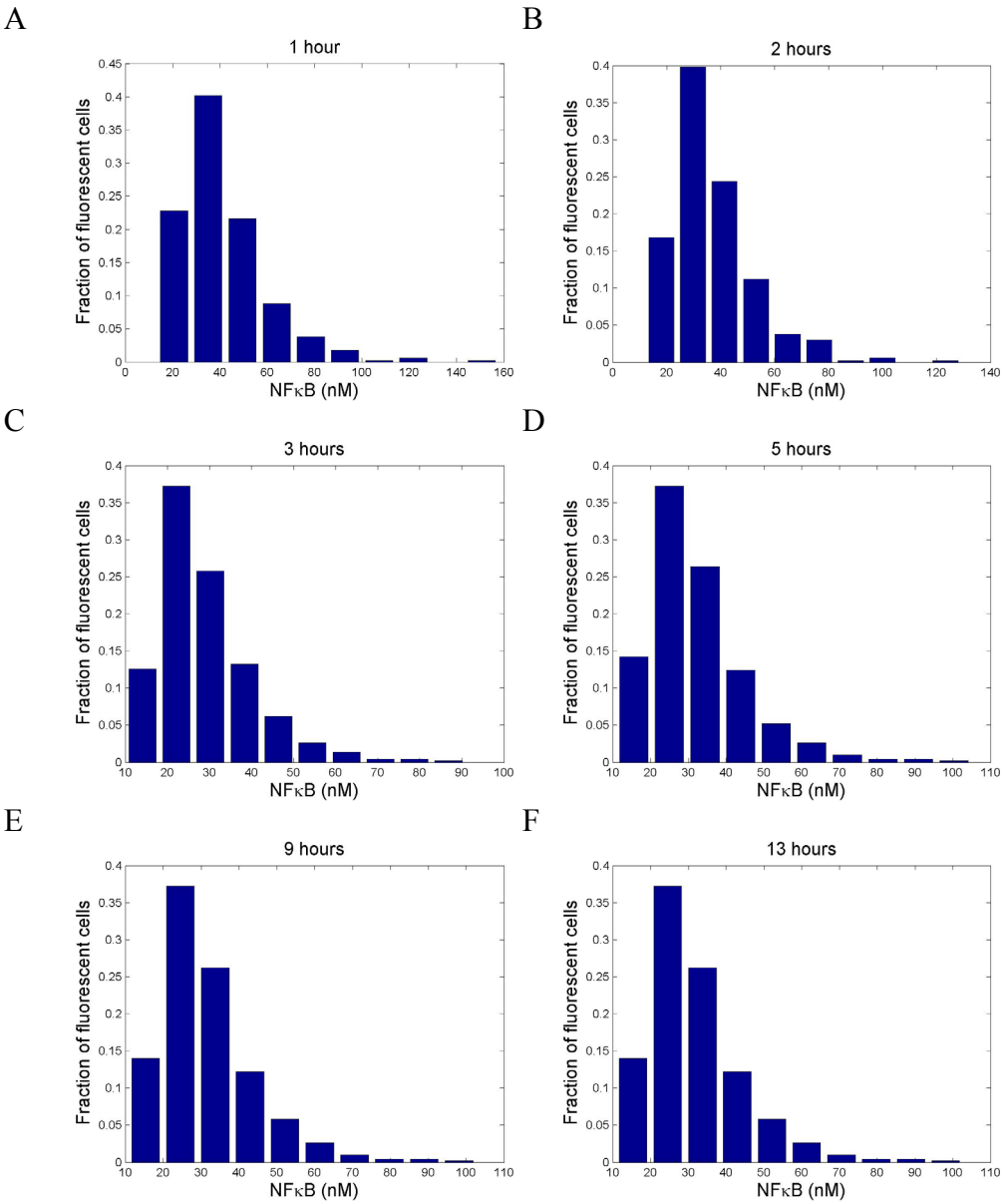


Figure 39. Distribution of NF-κB concentrations for H35 cells stimulated by 13 ng/ml of TNF-α at different points in time.

The distributions shown in Fig. 39 are derived from the data set from Fig. 38. Fig. 40 shows a comparison of the predicted fluorescence intensity distribution (the red curve) and the experimental data for one of the testing data sets. It can be seen that the inferred NF-κB distribution can induce a fluorescence intensity distribution that fits the

experimental data very well.

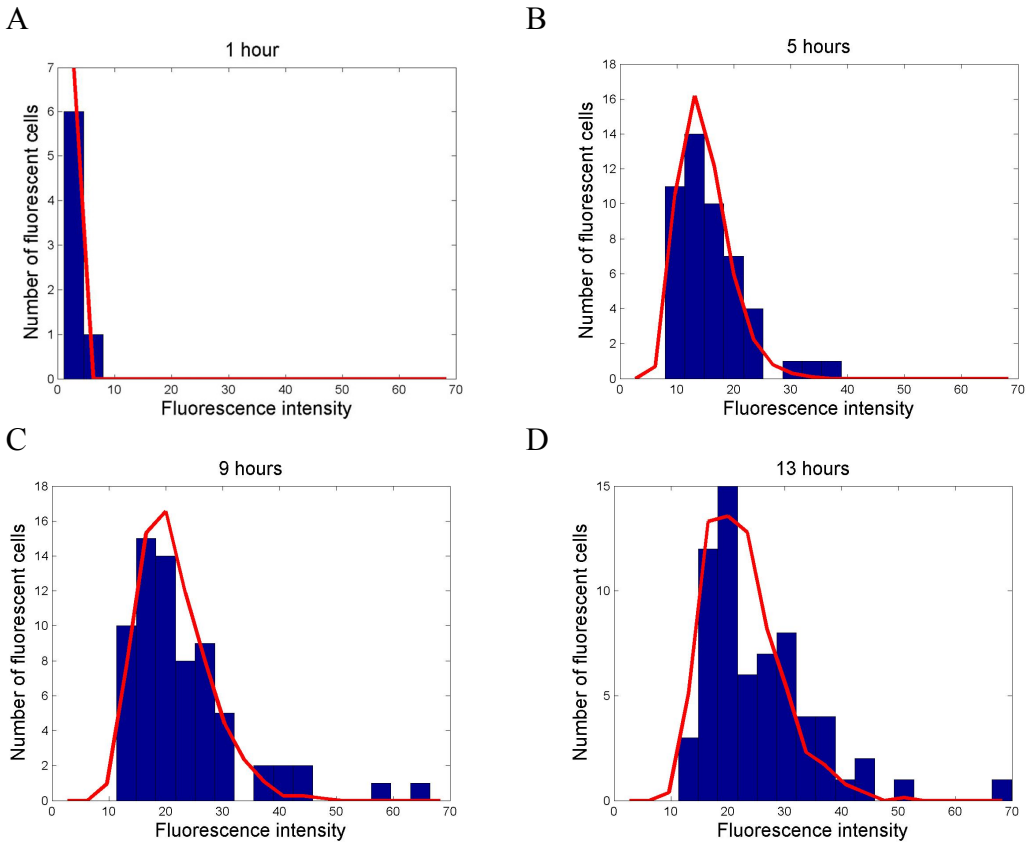


Figure 40. Comparison of experimental data and model prediction of the distribution of NF-κB at different points in time.

5.6 Summary

It is well established that transcription factors, which can be induced by a variety of stimuli, play a key role in signal transduction. However, even cells exposed to the same stimulus can exhibit a range of transcription factor activities due to stochasticity inherent in the process. This section addressed some aspects of this problem as it presented an approach for computing the transcription factor concentration distribution from fluorescence microscopy images of GFP reporter systems. This approach consists of

several steps that were individually introduced in this paper: detection of individual cells (instead of a collection of cells that border each other), determination of the distribution of the fluorescence intensity among a population of cells, and solution of an inverse problem that computes the transcription factor concentration distribution over time.

The presented approach was used to derive a distribution of the NF-κB concentrations from three sets of experimental data for H35 cells stimulated by 13 ng/ml TNF-α. The results showed that the presented image analysis was able to identify individual fluorescent cells from the fluorescence microscopy images, and the inferred NF-κB concentration distribution was able to provide a prediction of the fluorescence intensity distribution that was in good agreement with experimental data not used for deriving the model.

6. CONCLUSION

6.1 Conclusion and Contribution

Many systems investigated in Systems Biology are characterized by a large number of proteins and uncertain parameters, yet only a limited amount of quantitative data is generally available. The dissertation uses two different approaches to address this point. In the first approach, a model simplification procedure for signal transduction pathways is derived such that the most important verifiable relationships between concentrations of several different proteins are retained. In the second approach, quantitative measurement techniques for transcription factor concentration are developed. Based on the presented techniques, experimentally-verified models of signaling pathways involved in inflammation, such as the IL-6 signaling pathway and the TNF-α ~ NF-κB signaling pathway have been developed.

Section 3 presents an approach for simplifying signal transduction models such that 1) the model size is significantly reduced such that the model can be validated using available experimental data, and 2) the physical interpretation of the remaining states and parameters is retained. In the presented approach, sensitivity analysis is used to cluster reaction parameters with highly correlated effects, and observability analysis is performed to determine which states associated with each cluster of parameters are retained in the reduced model. The presented technique is used to derive a simplified version of an IL-6 signal transduction model. The number of equations and parameters in the model has been reduced from 65 to 13 and from 111 to 19, respectively. The

simplified model's ability to predict the dynamic behavior of the states for different conditions is evaluated at two different scenarios: predicting profiles of the states for a cell where SHP2 phosphorylation is blocked, predicting profiles for a cell where the value of the input and the values of the initial concentrations of the proteins vary randomly in a 30% range around the normal values. It is found that the simplified model is able to adequately describe the dynamics of some key proteins even if large uncertainty exists in the structure and parameter values of the signal transduction pathway model. It is also shown that the identifiability of the simplified model has been improved significantly. One reason for this is the correlation between parameters has been reduced via the presented procedure for selecting retained state variables for the simplified model from the states associated with different clusters of parameters.

Section 4 presents a method for quantitatively determining transcription factor concentration profiles from GFP reporter systems. An image analysis based on K-means clustering and principal component analysis is used to identify fluorescent cell regions in fluorescent images. Based on this, fluorescence intensity profiles are calculated from a time series of images. A model is developed to describe the dynamics of a reporter (i.e., green fluorescent reporter) concentration in response to a given transcription factor dynamics. Based on this model, a system inversion procedure has been developed to obtain transcription factor concentration profiles from fluorescence intensity profiles. The presented techniques are first applied to TNF-α ~ NF-κB signaling pathway and then to IL-6 signaling pathway. It is found that the image analysis method is able to detect fluorescent cell regions in fluorescent images correctly. The derived quantitative

data for transcription factor NF-κB is consistent with the available qualitative data in literature. The derived data for NF-κB is used to estimate the parameters of a developed model for TNF-α ~ NF-κB signaling pathway. It is found that the updated TNF-α model predicts experimental data very well. In this illustrative example, three potential profiles for transcription factors have been assumed for solving the inverse problem. This constraint can be relaxed by integrating signaling pathway models that describe the dynamics of transcription factors with the model describing GFP dynamics to estimate parameters in signaling pathway models. In this case, transcription factor profiles can be obtained from the model with the estimated parameters. This is illustrated by applying fluorescence intensity profiles from the GFP reporter systems for nuclear STAT3 and C/EBPβ to re-estimate the parameters of the simplified IL-6 model obtained in Section 3.

Section 5 extends the techniques presented in Section 4 to calculate transcription factor distribution profiles from GFP reporter data. Instead of focusing on average fluorescence intensity over fluorescent cell regions, this section develops an image analysis technique to identify individual fluorescent cells from fluorescent images and then obtains the distribution of fluorescence intensity at different points in time. An approach for solving an inverse problem is then presented to calculate the distribution of transcription factor concentrations from the distribution of fluorescence intensity. The presented techniques are then applied to the experiment data for TNF-α ~ NF-κB signaling pathway. It turns out that the image analysis method can identify individual fluorescent cells from fluorescent images correctly. Distribution of fluorescence intensity is obtained at different points in time. The distribution of NF-κB concentrations

is then obtained by solving the inverse problem. It is found that the inferred NF-κB concentration distribution is able to provide a prediction of the fluorescence intensity distribution that is in good agreement with experimental data not used for deriving the model.

The contribution made by this dissertation can be summarized as follows.

1) The presented model simplification approach can be applied to any ODE model of signal transduction pathways. This approach can significantly improve the identifiability of the model. Furthermore, the presented parameter clustering approach can be used to investigate the correlation between parameters in the model while the observability analysis approach can be used to select the appropriate proteins for measurement.
2) The integrated experimental and modeling approach for determining transcription factor profiles from fluorescent reporter data can be applied to any transcription factors if the corresponding GFP reporter systems can be developed and fluorescent images can be created. The quantitative data for the transcription factors NF-κB, C/EBPβ, and nuclear STAT3 derived in this dissertation can be used by other researchers for developing models where these transcription factors are involved. The data for these transcription factors can be provided upon request.
3) Two experimentally verified mathematical models have been developed for two different signal transduction pathways, i.e., the IL-6 signaling and the TNF-α ~ NF-κB signaling. These models can be integrated with the models for other

systemic inflammatory mediators, leading to an improved understanding of the molecular mechanisms involved in the acute phase response and thus an improved treatment of complications arising from inflammatory disorders. The MATLAB programs for these two models are available on the website (http://www.che.tamu.edu/orgs/groups/Hahn/Models/models.html).

4) Two image analysis algorithms are presented in this dissertation. The first one that only identifies fluorescent cell regions can be applied to any fluorescent images if information of individual fluorescent cells is not highly required and fluorescence intensity is the main useful information that can be extracted from images. In the case that information of individual cells or objects is important for investigation, the second image analysis approach can be applied as it can be used to identify individual objects from images but also detect edges between bordering objects. These two image analysis methods should be of interest for the pattern recognition community. The MATLAB program of a graphic user interface for the first image analysis method is available on the website (http://www.che.tamu.edu/orgs/groups/Hahn/Image_Analysis/index.html).

5) The developed model describing GFP dynamics can be used as a soft sensor to predict fluorescence dynamics in response to a given transcription factor dynamics. In addition, it can be used as the model for solving an inverse problem via which the information about the transcription factors is obtained from the fluorescence intensities. This model should be applicable to transcription factors other than NF-κB, C/EBPβ and nuclear STAT3. The MATLAB program for this

model can be provided upon requested.

6) The presented approach for solving an inverse problem can be used to derive the input profile from the output profile, even when only noisy and limited data are available for the output. Since quantitative data are difficult to obtain in Systems Biology and the data is usually of noise, this approach provides an applicable solution for this situation. In this approach, no prior knowledge of transcription factor dynamics is needed, although the prior knowledge can make the inverse problem being solved faster.

6.2 Suggestions for Further Work

Four suggestions are given below in each subsection for the possible extensions of this dissertation.

6.2.1 Development of a Cell-population Model for TNF-α ~ NF-κB Signaling Pathway

Distribution information of transcription factor NF-κB at different points in time has been obtained in Section 5. This information can be used to further develop a cell-population model for TNF-α ~ NF-κB signaling pathway. It is becoming increasingly evident that gene expression is stochastic, that is, there is significant variability between individual cells in the expression of different genes. Stochastic changes in transcription can arise due to changes in the levels of promoter binding activity of regulatory molecules controlling transcription (i.e., transcription factors), or randomness in the half-lives of the proteins involved in transcription (Tabor et al., 2008). This is the case for the gene expression of proteins A20 and IκBα in TNF-α ~ NF-κB signaling pathway, which resulting in the heterogeneity of NF-κB concentrations among different fluorescent cells

(Lipniacki et al., 2006).

One approach to model the distribution of NF-κB is using a stochastic switch to address the activity of A20 and IκBα in TNF-α ~ NF-κB signaling pathway, e.g., the status of the transcription of these two components, i.e., "ON" and "OFF", is described with probabilities determined by regulatory factors (Lipniacki et al., 2006). This is an ensemble approach as the stochasticity of NF-κB activation is counted from the multiple simulation results of the ODE model for a single cell. Another potential approach to address the stochasticity shown in NF-κB concentration distribution is specifying the stochasticity existing in the values of some reaction parameters. In this approach, a probability density function is assigned for each selected parameter, and the determining sets of parameters can then be used to describe the distribution of transcription factor activities over the population. Based on the ODE model for the single cell and probability density functions for the selected parameters, a cell-population model is described by partial differential equation (PDE) model. Compared to the ensemble approach, this approach provides a better way for theoretical analysis. However, the following issues need to be addressed for this approach.

First, a parameter selection approach needs to be developed for taking uncertainty in the model into account, as a parameter set which may be the best to estimate for one set of nominal values may not be the best set for other nominal values. It is important to note here that since the data set will consist of a distribution of transcription factor activity data, the parameter values will also follow a distribution which can result in different parameters being selected. As sensitivity vectors only correspond to the

nominal parameter values, the described technique needs to be relaxed to account for variations in the parameter values as they cannot be known prior to estimation.

Second, an approach to derive NF-κB concentration distribution from the PDE cell-population model should be investigated for specifying the probability density functions of the selected parameters. One potential solution to predict the output distribution from the distribution of the values of parameters is using a particle filtering approach presented in Rawlings and Bakshi, 2006. The predicted output distribution is then used to specify the probability density function of the selected parameters.

After solving the two issues mentioned above, a cell-population model should be developed to address the stochasticity shown in NF-κB concentration distribution. This approach can be implemented to IL-6 signaling pathway if distribution of nuclear STAT3 and C/EBPβ can be obtained, as the ODE model for a single cell for this signaling pathway is available.

6.2.2 An Alternative Experimental Approach to Monitor Fluorescence Dynamics

Although the present imaging system can monitor fluorescence dynamics well, the following issues need to be taken into account when an alternative experimental approach is investigated: (1) The fluorescence intensity data from the present imaging system is of a certain amount of noise. This can be concluded from the error-bar fluorescence intensity profiles shown in Fig. 20. The signal-noise ratio (SNR) for the present experiment data is around 36 dB. (2) The stimulation profile is not easy to change during experiment for the present imaging systems. For example, it is hard to change the concentration of the stimulating cytokine in the imaging period. Some

complicated stimulation patterns such as multiple pulse stimulation with different frequencies or durations are impossible to implement at this stage. (3) the present imaging system doesn't allow to take fluorescent images at a high frequency pace as the cells are exposed to UV light during the experiment and cells are deactivated or even dying for high-frequency exposure to UV light. This leads to a low frequency sampling pace and thus results in fluorescence intensity profiles sampled at only limited points in time. This makes it challenging to infer the transcription factor profiles from the fluorescence intensity profile by solving an inverse problem. A conservative approach (i.e., parametric approach) is used in this dissertation and three potential profiles are assumed for transcription factors to solve the inverse problem. Non-parametric approach is applicable to the inverse problem if fluorescence intensity is sampled at more time points and the noise from the imaging systems is reduced to some extent. (4) The fluorescent cells in the fluorescent images of GFP reporter systems tends to border each other. This makes the task to identify individual fluorescent cells challenging. Although the program presented in Section 5 can be applied to identify individual fluorescent cells, it is time-consuming as it is not that easy to separate the cells bordering each other and one experiment dataset generally has thousands of images. If the fluorescent cells can be separated experimentally, this alleviates the computational load for image analysis programming. In addition, this might result in a more accurate identification of individual fluorescent cells. (5) After the individual fluorescent cells are identified (no matter via image analysis program or experimental approaches), the effect of cell population on signal transduction can be investigated if the experiment approach can

control cell density. It is an interesting topic for investigation as it is found that cellular phenotypes can be significantly altered in the absence of proximity between similar cells and signals (Losick and Desplan, 2000).

Based on the above discussion, an alternative experiment approach that can reduce the noise of image, allow time-varying stimulation patterns, enable sampling fluorescent images at a high frequency pace, separate the fluorescent cells, and control cell population density during the experiment, is highly required. In the author's opinion, improving the quality of images via an alternative experimental approach is a better way than developing improved image analysis programs to extract useful information from the noisy and limited fluorescent images. Using a microfluidic device for generating the experimental data is a promising approach to address some of the points mentioned above as it is found that this approach can create a sufficiently large data set for data analysis and that it can separate cells by specifically positioning cells at different locations in the microfluidic chamber (Bhatia et al., 1997; King et al., 2007; King et al., 2008; Snykers et al., 2006; Thompson et al., 2004, Wieder et al., 2005). In addition, it is convenient to change the stimulation patterns in the microfluidic device. Therefore, microfluidic devices for monitoring fluorescence dynamics for specific transcription factors need to be designed, and the micropatterning techniques for positioning cells are also good subjects for investigation.

6.2.3 Investigation of IL-6 and IL-10 Signaling in Steatosis via Mathematical Modeling

As mentioned in Section 1, IL-6 signaling plays an important in the progression of steatosis, as the two transcription factors involved in IL-6 signaling, i.e. nuclear STAT3

and C/EBPβ, are found to have opposite effect on the progression of non-alcoholic fatty liver disease. In addition to IL-6, IL-10 also plays an important role in regulating the activation level of STAT3. As IL-10 signaling utilizes only the Jak-STAT pathway (Murray, 2007), it can lead to STAT3 activation without activating C/EBPβ. When a combination of IL-6 and IL-10 are applied, binding of IL-10 to its receptor leads to activation of Jak and STAT3 and the synthesis of SOCS3. The IL-10 induced SOCS3 then inhibits binding of Jak to the IL-6 receptor but not to the IL-10 receptor; thereby, selectively inhibits IL-6 signaling (Murray, 2006; Murray, 2007). Since no C/EBPβ is activated by IL-10, C/EBPβ activation should be inhibited when IL-6 signaling is inhibited; however, this has not yet been shown in hepatocytes. A result of this would be that IL-10 in conjunction with IL-6 should be able to independently regulate the activation levels of STAT3 and C/EBPβ. While the dynamic behavior of a few individual molecules, e.g., STAT3 and SOCS3, in the IL-6 pathway is known (Huang et al., 2007; Schoeberl, et al., 2002; Singh, et al., 2006; Yamada, et al., 2003), comprehensive signaling pathway dynamics and interactions for IL-6 and IL-10 signal transduction, as well as their impact on disease progression, are not well understood. In order to address this issue, a model for IL-6 and IL-10 signal transduction should be developed. Moya et al, 2010, has developed a preliminary model for IL-6 and IL-10 signaling. However, the model parameters involved in IL-10 signaling pathway need to be further verified by experimental data. For this purpose, a GFP-reporter system for nuclear STAT3 stimulated by IL-10 needs to be developed. The techniques for deriving quantitative data for transcription factors presented in this dissertation can then be applied to the obtained

GFP reporter data for IL-10 signaling. In addition, the interaction between IL-6 signaling and IL-10 signaling can be experimentally investigated if the GFP reporter system for IL-10 signaling can be integrated with those for IL-6 signaling. In this case, the experiment system for monitoring the dynamics of multiple transcription factors stimulated by multiple cytokines needs to be proposed.

6.2.4 Development of a Comprehensive Model for the Signaling Pathways Involved in Acute Phase Response

In addition to IL-6 and TNF-α, IL-1 and OSM are the other two systemic inflammatory mediators. Besides the signaling pathways shown in Roth et al, 2001 that are involved in acute phase response, TNF-α signaling is taken as another signal transduction pathway for acute phase response in Fig. 41.

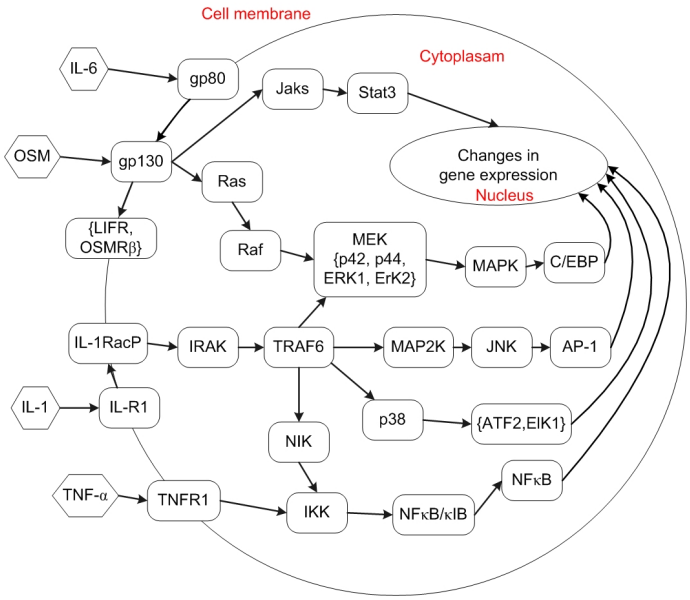


Figure 41. Overview of the signaling pathways involved in acute phase response. The crosstalk between IL-6 signaling, OSM signaling, IL-1 signaling, and TNF-α signaling determines the changes in gene expression.

It can be seen from Fig. 41 that these signaling pathways have strong interaction. Specifically, IL-6 signaling and OSM signaling share receptor gp130 and the pathway associated with this receptor, IL-6 and IL-1 signaling share parts of the Erk-C/EBPβ pathway, and TNF-α and IL-1 signaling share parts of NF-κB pathway. The crosstalk between these signaling pathways determines the change of gene expression and thus the function of the cell. Since each of these signaling pathways is complicated for its large number of components and the crosstalk of its components, mathematical models should be developed for each of these signaling pathways. Based on these models, a comprehensive mathematical model for the regulatory mechanism underlying acute phase response can be developed. Although mathematical models are available for IL-6 signaling and TNF-α signaling (Huang et al., 2008; Huang et al., 2010; Singh et al, 2006), models for IL-1 signaling still needs to be developed. IL-1 signaling consists of four pathways, i.e., Erk-C/EBPβ pathway, JNK pathway, p38 pathway, and NF-κB pathway. Mathematical models for Erk-C/EBPβ pathway and NF-κB pathway can be developed if the model for the signaling pathway from IL-1 to TRAF6 is derived, as the models from TRAF to C/EBPβ and to NF-κB can be respectively taken from the developed models for IL-6 signaling and TNF-α signaling. In order to develop model for the pathway from IL-1 to TRAF6 and those pathways following TRAF6 such as TRAF6-JNK and TRAF6-p38, a detailed literature review needs to be implemented. Furthermore, the available models in the database mentioned in Section 1 should be also investigated. Based on this, ODE model for IL-1 signaling should be derived. To estimate the reaction parameters involved in the model, experimental data for the

transcription factors involved in IL-1signaling should be obtained. The techniques presented in Section 4 and 5 can be applied if the corresponding GFP reporter systems for those transcription factors are developed. Compared with the model for IL-1 signaling, the model for OSM signaling is easier to develop as it can share the model with IL-6 signaling for the pathway after receptor gp130. After an experimentally verified model has been developed for IL-1 signaling and OSM signaling, these models are then integrated with the models for IL-6 signaling and TNF-α signaling. The resulting integrated model can be used to study the crosstalk between these signaling pathways, investigate the regulatory mechanism underlying acute phase response, and detect some biomarkers for drug development to improve the treatment of inflammatory disorders.

REFERENCES

Abiru S, Migita K, Maeda Y, Daikoku M, Ito M, Ohata K, Nagaoka S, Matsumoto T, Takii Y, Kusumoto K, Nakamura M, Komori A, Yano K, Yatsuhashi H, Eguchi K, Ishibashi H. 2006. Serum cytokine and soluble cytokine receptor levels in patients with non-alcoholic steatohepatitis. Liver Int 26:39–45.

Aldridge BB, Saez-Rodriguez J, Muhlich JL, Sorger PK, Lauffenburger DA. 2009. Fuzzy logic analysis of kinase pathway crosstalk in TNF/EGF/insulin-induced signaling. PLoS Comput Biol 5(4):e1000340.

Androulakis IP. 2000. Kinetic mechanism reduction based on an integer programming approach. AIChE J 46:361–371.

Assmus HE, Boldt S, Wolkenhauer O. 2009. Reverse engineering of biological networks. Methods in bioengineering: systems analysis of biological networks. Boston: Artech House 33–56.

Bailey JE, Ollis DF. 1986. Biochemical engineering fundamentals. New York: McGraw-Hill.

Bandhyopadhyay S, Soto-Nieves N, Macián F. 2007. Transcriptional regulation of T-cell tolerance. Semin Immunol 19(3):180–187.

Bao P, Zhang L, Wu X. 2005. Canny edge detection enhancement by scale multiplication. IEEE Trans Pattern Anal Mach Intell 27(9):1485–1490.

Benyon PR. 1979. The inversion of dynamic systems. Math Comput Simul XXI:335–339.

Bharati MHM, Macgregor JF. 1998. Multivariate image analysis for real–time process monitoring and control. Ind Eng Chem Res 37:4715–4724.

Bhatia SN, Yarmush ML, Toner M. 1997. Controlling cell interactions by micropatterning in co-cultures: hepatocytes and 3T3 fibroblasts. J Biomed Mater Res 34:89–99.

Brewer J, Huang Z, Singh AK, Misra M, Hahn J. 2007. Sensor network design via observability analysis and principal component analysis. Ind Eng Chem Res 46(24):8026–8032.

Brockett RW. 1970. Finite dimensional linear systems. New York: Wiley.

Birtwistle MR, Hatakeyama M, Yumoto N, Ogunnaike BA, Hoek JB, Kholodenko BN. 2007. Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. Mol Syst Biol 3(144):1–16.

Brightman FA, Fell DA. 2000. Differential feedback regulation of the MAPK cascade underlies the quantitative differences in EGF and NGF signalling in PC12 cells. FEBS Lett 482:169–174.

Brown AJ. 1902. Enzyme action. J Chem Soc 81:373–386.

Canny J. 1986. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 8(6):679–698.

Chang R, Brauer W, Stetter M. 2008. Modeling semantics of inconsistent qualitative knowledge for quantitative Bayesian network inference. Neural Networks 21:182–192.

Clark JM. 2006. The epidemiology of nonalcoholic fatty liver disease in adults. J Clin Gastroenterol 40: S5–10.

Chaudhuri S, Chatterjee S, Katz N, Nelson M, Goldbaum M. 1989. Detection of blood vessels in retinal images using two-dimensional matched filters. IEEE Trans Pattern Anal Mach Intell 8(3):263–269.

Corvinus FM, Orth C, Morigg R, Tsareva SA, Wagner S, Pfitzner EB, Baus D, Kaufmann R, Huberb LA, Zatloukal K, Beug H, Ohlschlager P, Schutz A, Halbhuber KJ, Friedrich K. 2005. Persistent STAT3 activation in colon cancer is associated with enhanced cell proliferation and tumor growth. Neoplasia 7(6):545–555.

Cukier RI, Fortuin CM, Shuler KE, Petschek AG, Schaibly JH. 1973. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients – I. Theory. J Chem Phys 59(8):3873–3878.

Dano S, Madsen MF, Schmidt H, Cedersund G. 2006. Reduction of a biochemical model with preservation of its basic dynamic properties. FEBS J 273(21):4862–4877.

Davidov EJ, Holland JM, Marple EW, Naylor S. 2003. Advancing drug discovery through systems biology. Drug Discovery Today 8(4):175–183.

Deimund M, Huang Z, Hahn J. 2010. A graphic user interface of image analysis program for green fluorescent images. http://www.che.tamu.edu/orgs/groups/Hahn/Image_Analysis/index.html.

Diehl AM. 2004. Tumor necrosis factor and its potential role in insulin resistance and nonalcoholic fatty liver disease. Clin Liv Dis 8:619–638.

Ding C, He X. 2004. K-means clustering via principal component analysis. Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 225–232.

Dokoumetzidis A, Aarons L. 2009. Proper lumping in systems biology models. IET Syst Biol 3(1):40–51.

Dudley DG. 1985. Comments on SEM and the parametric inverse problem. IEEE Trans Antennas Propag AP-33(1):119–120.

Edwards K, Edgar T, Manousiouthakis V. 1998. Kinetic model reduction using genetic algorithms. Comput Chem Eng 22:239–246.

Efroni S, Schaefer CF, Buetow KH. 2007. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. PLoS ONE 2(5):e425.

Elnitski L, Jin VX, Farnham PJ, Jones SJ. 2006. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res 16:1455–1464.

Eungdamrong NJ, Iyengar R. 2004. Modeling cell signaling networks. Biol Cell 96:355–362.

Fasshauer M, Kralisch S, Klier M, Lossner U, Bluher M, Klein J, Paschke R. 2004. Insulin resistance-inducing cytokines differentially regulate SOCS mRNA expression via growth factor- and JAK/Stat-signaling pathways in 3T3-L1 adipocytes. J Endocrinol 181:129–138.

Feldmann U, Hasler M, Schwarz W. 1998. Communication by chaotic signals: the inverse system approach. Int J Circuit Theory Appl 24(5): 551–579.

Fischer P, Hilfiker-Kleiner D. 2008. Role of gp130-mediated signaling pathways in the heart and its impact on potential therapeutic aspects. Br J Pharmacol 153:S414–S427.

Frank PM. 1978. Introduction to system sensitivity theory. New York: Academic Press.

Friedman N, Linial M, Nachman I, Pe'er D. 2000. Using Bayesian networks to analyze expression data. J Comput Biol 7:601–620.

Geback T, Koumoutsakos P. 2009. Edge detection in microscopy images using curvelets. BMC Biol 10:75.

Geladi P, Grahn H. 1996. Multivariate image analysis. Chichester, UK: John Wiley & Sons Press.

Guimerà R, Nunes Amaral LA. 2005. Functional cartography of complex metabolic networks. Nature 433:895–900.

Hahn J, Edgar TF. 2001. A gramian based approach to nonlinearity quantification and model classification. Ind Eng Chem Res 40:5724–5731.

Hahn J, Edgar TF. 2002. An improved method for nonlinear model reduction using balancing of empirical gramians. Comput Chem Eng 26:1379 –1397.

Hahn J, Edgar TF, Marquardt W. 2003. Controllability and observability covariance matrices for the analysis and order reduction of stable nonlinear systems. J Process Control 13:115–127.

Haralick RM. 1984. Digital step edges from zero crossing of second directional derivatives. IEEE Trans Pattern Anal Mach Intell 6(1):58–68.

Haralick RM, Sternberg SR, Zhuang X. 1987. Image analysis using mathematical morphology. IEEE Trans Pattern Anal Mach Intell 9 (4):532–550.

Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. 2001. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pac Symp Biocomput 1:422–433.

Hasty J, McMillen D, Isaacs F, Collins JJ. 2001. Computational studies of gene regulatory networks: in numero molecular biology. Nat Rev Genet 2:268–279.

Heinrich PC, Behrmann I, Muller-Newen G, Schaper F, Graeve L. 1998. Interleukin-6-type cytokine signalling through the gp130/Jak/STAT pathway. Biochem J 334(Pt 2): 297–314.

Heinrich PC, Behrmann I, Haan S, Hermanns HM, Muller-Newen G, Schaper F. 2003. Principles of interleukin (IL)-6-type cytokine signaling and its regulation. Biochem J 374:1–20.

Henson MA, Seborg DE. 1996. Nonlinear process control. Upper Saddle River, NJ: Prentice Hall.

Hoffmann A, Levchenko A, Scott ML, Baltimore D. 2002. The IkB–NF-κB signaling module: temporal control and selective gene activation. Science 298(8):1241–1245.

Hoffmann A, Natoli G, Ghosh G. 2007. Transcriptional regulation via the NF-kappaB signaling module. Oncogene 25(51):6706–6716.

Hong F, Radaeva S, Pan HN, Tian Z, Veech R, Gao B. 2004. Interleukin 6 alleviates hepatic steatosis and ischemia/reperfusion injury in mice with fatty liver disease. Hepatology 40:933–941.

Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. J of Educ Psychol 24:417–441.

Hsiao Y, Chuang C, Lu Y, Jiang J. 2006. Robust multiple objects tracking using image segmentation and trajectory estimation scheme in video frames. Image Vision Comput 24(10):1123–1136.

Huang Z, Chu Y, Senocak F, Jayaraman A., Hahn J. 2007. Model update of signal transduction pathways in hepatocytes based upon sensitivity analysis. Proceedings of Foundations of Systems Biology 2007, Stuttgart, Germany, 45-50.

Huang Z, Senocak F, Jayaraman A, Hahn J. 2008. Integrated modeling and experimental approach for determining transcription factor concentrations from fluorescent reporter profiles. BMC Syst Biol 2:64.

Huang Z, Moya C, Cheng P, Jayaraman A, Hahn J. 2009a. In silico investigation of IL-6 and IL-10 signaling in steatosis. Proceedings of Foundations of Systems Biology in Engineering 2009, Denver, CO, 28–31.

Huang Z, Hahn J. 2009b. Fuzzy modeling of signal transduction networks. Chem Eng Sci 64(9):2044–2056.

Huang Z, Chu Y, Hahn J. 2010. Model simplification procedure for signal transduction pathway models: an application to IL-6 signaling. Chem Eng Sci 65(6):1964–1975.

Inoue H, Ogawa W, Ozaki M, Haga S, Matsumoto M, Furukawa K, Hashimoto N, Kido Y, Mori T, Sakaue H, Teshigawara K, Jin S, Iguchi H, Hiramatsu R, LeRoith D, Takeda K, Akira S, Kasuga M. 2004. Role of STAT-3 in regulation of hepatic gluconeogenic genes and carbohydrate metabolism in vivo. Nat Medicine 10:168–174.

Ideker T, Lauffenburger D. 2003. Building with a scaffold: emerging strategies for high- to low-level cellular modeling. Trends Biotechnol 21(6):255–262.

Jackson JE. 2003. A user's guide to principal components. Hoboken, NJ: John Wiley & Sons Press.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. 2000. The large-scale organization of metabolic networks. Nature 407:651–654.

Judd LM, Alderman BM, Howlett M, Shulkes A, Dow C, Moverley J, Grail D, Jenkins BJ, Ernst M, Giraud AS. 2004. Gastric cancer development in mice lacking the SHP2 binding site on the IL-6 family co-receptor gp130. Gastroenterology 126(1):196–207.

Karlebach G, Shamir R. 2008. Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol 9:770–780.

Kauffman SA. 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. J Theoret Biol 22:437–467.

Kaufman L, Rousseeuw PJ. 1990. Finding groups in data: an introduction to cluster analysis, New York: John Wiley & Sons Press.

Kholodenko BN, Demin OV, Moehren G, Hoek JB. 1999. Quantification of short term signaling by the epidermal growth factor receptor. J Biol Chem 274:30169–30181.

Kholodenko BN. 2006. Cell-signalling dynamics in time and space. Nat Rev Mol Cell Biol 7:165–176.

King KR, Wang S, Jayaraman A, Toner M, Yarmush ML. 2007. A highthroughput microfluidic real-time gene expression living cell array. Lab-on-Chip 7:77–85.

King KR, Wang S, Jayaraman A, Yarmush ML, Toner M. 2008. Microfluidic flow-encoded switching for parallel control of dynamic cellular microenvironments. Lab-on-Chip 8:107–116.

Klamt S, Saez-Rodriguez J, Lindquist JA, Simeoni L, Gilles ED. 2006. A methodology for the structural and functional analysis of signalling and regulatory networks. BMC Bioinf 7:56.

Klamt S, Saez-Rodriguez J, Gilles E. 2007. Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC Syst Biol 1:2.

Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H. 2005. Systems biology in practice: concepts, implementation and application. Weinheim, Germany: Wiley.

Kruger R, Heinrich R. 2004. Model reduction and analysis of robustness for the Wnt/β-catenin signal transduction pathway. Genome Inf 15(1):138–148.

Kurien BT, Scofield RH. 2006. Western blotting. Methods 38(4):283–293.

Lalor PF, Faint J, Aarbodem Y, Hubscher SG, Adams DH. 2007. The role of cytokines and chemokines in the development of steatohepatitis. Semin Liver Dis 27:173–193.

Lang R, Pauleau AL, Parganas E, Takahashi Y, Mages J, Ihle JN, Rutschman R, Murray PJ. 2003. SOCS3 regulates the plasticity of gp130 signaling. Nat Immunol 4:546–550.

Lee EG, Boone DL, Chai S, Libby SL, Chien M, Lodolce JP, Ma A. 2000. Failure to regulate TNF-induced NF-κB and cell death responses in A20-deficient mice. Science 289:2350–2354.

Li S, Assmann S, Albert R. 2006. Predicting essential components of signal transduction networks: dynamic model of guard cell abscisic acid signaling. PLoS Biol 10:e312.

Lia T, Wang S, Zhao N. 2009. Gray-scale edge detection for gastric tumor pathologic cell images by morphological analysis. Comput Biol Med 39(11):947–952.

Liebermeister W, Baur U, Klipp E. 2005. Biochemical network models simplified by balanced truncation. FEBS J 272:4034–4043.

Lin F, Olbrot AW. 1996. An LQR approach to robust control of linear systems with uncertain parameters. Proceedings of the 35th IEEE Conference on Decision and Control. Kobe, Japan, 4:4158–4163.

Lipniacki T, Paszek P, Brasier AR, Luxon B, Kimmel M. 2004. Mathematical model of NF-κB regulatory module. J Theor Biol 228:195–215.

Lipniacki T, Paszek P, Brasier AR, Luxon B, Kimmel M. 2006. Stochastic regulation in early immune response, Biophys J 90:725–742.

Lloyd SP. 1982. Least squares quantization in PCM. IEEE Trans on Inf Theory 28(2):129–137.

Losick R, Desplan C. 2008. Stochasticity and cell fate. Science 320(5872):65–68.

Lucas JF. 2005. Bayesian network modelling through qualitative patterns. Artif Int 163:233–263.

Mamdani EH, Assilian S. 1975. An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies 7(1):1–13.

Markusson O. 2002. Model and system inversion with applications in nonlinear system identification and control. PhD Dissertation, Royal Institute of Technology, Stockholm, Sweden.

McRae GJ, Tilden JW, Seinfeld JH. 1982. Global sensitivity analysis-a computational implementation of the Fourier amplitude sensitivity test (FAST). Comput Chem Eng 6(1):15–25.

Marr D, Hildreth EC. 1980. Theory of edge detection. Proc R Soc B-207:187–217.

McClain CJ, Mokshagundam SP, Barve SS, Song Z, Hill DB, Chen T, Deaciuc I. 2004. Mechanisms of non-alcoholic steatohepatitis. Alcohol 34:67–79.

Méndez-Sánchez N, Arrese M, Zamora-Valdés D, Uribe M. 2007. Current concepts in the pathogenesis of nonalcoholic fatty liver disease. Liver Int 27:423–433.

Meyer F. 1979. Iterative image transformation for an automatic screening of cervical smears. The Journal of Histochemistry and Cytochemistry 27(1):128–135.

Moore BC. 1981. Principal component analysis in linear systems: Controllability, observability, and model reduction. IEEE Trans Auto Control 26(1):17–32.

Moya C, Huang Z, Cheng P, Jayaraman A, Hahn J. 2010. Investigation of IL-6 and IL-10 signaling in steatosis via mathematical modeling. Submitted to IET Syst Biol.

Murray PJ. 2006. Understanding and exploiting the endogenous interleukin-10 /STAT3-mediated anti-inflammatory response. Curr Opin Pharmacol 6:379–386.

Murray PJ. 2007. The JAK-STAT signaling pathway: Input and output integration. J Immunol 178:2623–2629.

Oda K, Matsuoka Y, Funahashi A, Kitano H. 2005. A comprehensive pathway map of epidermal growth factor receptor signaling. Mol Syst Biol 1: 2005.0010 .

Orton RJ, Sturm OE, Vyshemirsky V. 2005. Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway. Biochem J 392:249–261.

Pan Q, Saltzman AL, Kim YK, Misquitta C, Shai O, Maquat LE, Frey BJ, Blencowe BJ. 2006. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev 20(2):153–158.

Paulsson J. 2004. Summing up the noise in gene networks. Nature 427:415–418.

Perona P, Malik J. 1990. Scale-space and edge detection using anisotropic diffusion. IEEE Trans Pattern Anal Mach Intell 12(7):629–639.

Peters RA. 1995. A new algorithm for image noise reduction using mathematical morphology. IEEE Trans Image Process 4(5):554–568.

Petzold LR, Zhu W. 1999. Model reduction for chemical kinetics: an optimization approach. AIChE J 45:869–886.

Piazzi A, Visioli A. 2000. Minimum-time system-inversion-based motion planning for residual vibration reduction. IEEE/ASME Trans Mechatron 5(1):12–22.

Pokholok DK, Zeitlinger J, Hannett NM, Reynolds DB, Young RA. 2006. Activated signal transduction kinases frequently occupy target genes. Science 313:533–536.

Poon SS, Martens UM, Ward RK, Lansdorp PM. 1999. Telomere length measurements using digital fluorescence microscopy. Cytometry 36:267–278.

Prewitt JMS. 1970. Object enhancement and extraction. In: Lipkin BS, Rosenfeld A, editors. Picture analysis and psychopictorics. New York: Academic Press 75–149.

Puebla H, Alvarez-Ramirez J. 2001. Stability of inverse-system approaches in coherent chaotic communication. IEEE Trans Circuits Syst.-I: Fundamental Theory and Application 48(12): 1413–1423.

Purutcuoglu V, Wity E. 2008. Bayesian inference for the MAPK/ERK pathway by considering the dependency of the kinetic parameters. Bayesian Analysis 3(4):851–886.

Rahman SM, Schroeder-Gloeckler JM, Janssen RC, Jiang H, Qadri I, Maclean KN, Friedman JE. 2007. CCAAT/enhancing binding protein beta deletion in mice attenuates inflammation, endoplasmic reticulum stress, and lipid accumulation in diet-induced nonalcoholic steatohepatitis. Hepatoclogy 45:1108–1117.

Rangamani P, Sirovich L. 2007. Survival and apoptotic pathways initiated by TNF-a: modeling and predictions. Biotechnol Bioeng 97(5):1216–1229.

Raser JM, O'Shea EK. 2004. Control of stochasticity in eukaryotic gene expression. Science 304:1181–1184.

Rawlings JB, Bakshi BR. 2006. Particle filtering and moving horizon estimation. Comput Chem Eng 30:1529–1541.

Roberts LG. 1965. Machine perception of three-dimensional solids. Optical and electrooptical information processing. Cambridge, MA: MIT Press.

Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB. 2005. Gene regulation at the single-cell level. Science 307:1962–1965.

Roth CM, Kohen RL, Walton SP, Yarmush ML. 2001. Coupling of inflammatory cytokine signaling pathways probed by measurements of extracellular acidification rate. Biophys Chem 89:1–12.

Sabin J, Gray R. 1986. Global convergence and empirical consistency of the generalized Lloyd algorithm. IEEE Trans on Inf Theory 32(2):148–155.

Sachs K, Gifford D, Jaakkola T, Sorger P, Lauffenburger DA. 2002. Bayesian network approach to cell signaling pathway modeling. Science's STKE 148:E38.

Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. Science 308:523–529.

Saez-Rodriguez J, Simeoni L, Lindquist JA, Hemenway R, Bommhardt U, Arndt B, Haus UU, Weismantel R, Gilles ED, Klamt S, Schraven B. 2007. A logical model provides insights into T cell receptor signaling. PLoS Comput Biol 3(8):e163.

Sanyal AJ. 2005. Mechanisms of disease: pathogenesis of nonalcoholic fatty liver disease. Nature Clin Pract 2:46–53.

Schoeberl B, Eichler-Jonsson C, Gilles ED, Muller G. 2002. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nat Biotechnol 20:370–375.

Schuster S, Fell DA, Dandekar T. 2000. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat Biotechnol 18:326–332.

Sengupta TK, Talbot ES, Scherle PA, Ivashkiv LB. 1998. Rapid inhibition of interleukin-6 signaling and Stat3 activation mediated by mitogen-activated protein kinases. Proc Natl Acad Sci USA 95:11107–11112.

Serra J. 1982. Image analysis and mathematical morphology. London: Academic Press.

Shmulevich I, Dougherty ER, Kim S, Zhang W. 2002. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 18:261–274.

Singh AK, Hahn J. 2005. Determining optimal sensor locations for state and parameter estimation for stable nonlinear systems. Ind Eng Chem Res 44(15):5645–5659.

Singh AK, Hahn J. 2006. Sensor location for stable nonlinear dynamic systems: multiple sensor case. Ind Eng Chem Res 45(10):3615-3623.

Singh AK, Jayaraman A, Hahn J. 2006. Modeling regulatory mechanisms in IL-6 signal transduction in hepatocytes. Biotechnol Bioeng 95(5):850–862.

Sirdeshpande AR, Ierapetritou MG, Androulakis IP. 2001. Design of flexible reduced kinetic mechanisms. AIChE J 47:2461–2473.

Skogestad S, Postlethwaite I. 1997. Multivariable feedback control. New York: John Wiley & Sons.

Smits KW, Caroline EC, Kim SA, Sierd B, Oscar KP, Leendert HW. 2005. Stripping bacillus: comK auto-stimulation is responsible for the bistable response in competence development. Mol Microbiol 56(3):604–614.

Snykers S, Vanhaecke T, Papeleu P, Luttun A, Jiang Y, Vander Heyden Y, Verfaillie C, Rogiers V. 2006. Sequential exposure to cytokines reflecting embryogenesis: the key for in vitro differentiation of adult bone marrow stem cells into functional hepatocyte-like cells. Toxicol Sci 94(2):330–41.

Sobel I. 1978. Neighborhood coding of binary images for fast contour following and general array binary processing. Computer Graphics and Image Processing 8:127–135.

Spetsieris K, Zygourakis K, Mantzaris NV. 2009. A novel assay based on fluorescence microscopy and image processing for determining phenotypic distributions of rod-Shaped bacteria. Biotechnol Bioeng 102(2):598–615.

Steven F, Huang Z, Hahn J. 2010. Updated IL-6 signal transduction model. http://www.che.tamu.edu/orgs/groups/Hahn/Models/models.html.

Subramanian S, Srienc F. 1996. Quantitative analysis of transient gene expression in mammalian cells using the green fluorescent protein. J Biotechnol 49:137–151.

Sugeno M. 1985. Industrial applications of fuzzy control. New York: Elsevier Science Inc.

Sun C, Hahn J. 2006a. Parameter reduction for stable dynamical systems based on Hankel singular. Chem Eng Sci 61:5393–5403.

Sun C, Hahn J. 2006b. Nonlinear model reduction routines for MATLAB. http://che.tamu.edu/orgs/groups/Hahn/Model_Reduction/index.html.

Sun Z, Tsao T. 1999. Adaptive tracking control by system inversion. Proceedings of the 1999 American Control Conference, San Diego, CA, 1:29–33.

Tabor JJ, Bayer TS, Simpson ZB, Levy M, Ellington AD. 2008. Engineering stochasticity in gene expression. Mol Biosyst 4(7):754–761.

Taylor SR, Doyle FJ, Petzold LR. 2008. Oscillator model reduction preserving the phase response: application to the circadian clock. Biophys J 95:1658–1673.

Thompson DM, King KR, Wieder KJ, Toner M, Yarmush ML, Jayaraman A. 2004. Dynamic gene expression profiling using a microfabricated living cell array. Anal Chem 2004 76:4098–4103.

Thakar J, Pilione M, Kirimanjeswara G, Harvill E, Albert R. 2007. Modeling systems-level regulation of host immune responses. PLoS Comput Biol 6:e109.

Tomovic R, Vukobratovic M. 1972. General sensitivity theory. New York: Elsevier.

Torre V, Poggio TA. 1979. On edge diction. IEEE Trans Pattern Anal Mach Intell PAMI-8:147–163.

Ueki K, Kondo T, Tseng YH, Kahn CR. 2004. Central role of suppressors of cytokine signaling proteins in hepatic steatosis, insulin resistance, and the metabolic syndrome in the mouse. Proc Natl Acad Sci USA 101:10422–10427.

Vollmera U, Raisch J. 2006. Control of batch crystallization-A system inversion approach. Chem Eng Process 45(10):874–885.

Venkataraman S, Morrell-Falvey JL, Doktycz MJ, Qi H. 2005. Automated image analysis of fluorescence microscopic images to identify protein-protein interactions. Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Shanghai, China, 797-800.

Videla LA, Rodrigo R, Araya J, Poniachik J. 2006. Insulin resistance and oxidative stress interdependency in non-alcoholic fatty liver disease. Trends Mol Med 12:555–558.

Wieder KJ, King KR, Thompson DM, Zia C, Yarmush ML, Jayaraman A. 2005. Optimization of reporter cells for expression profiling in a microfluidic device. Biomed Microdevices 7:213–222.

Willner IR, Waters B, Patil SR, Reuben A, Morelli J, Riely CA. 2001. Ninety patients with nonalcoholic steatohepatitis: insulin resistance, familial tendency, and severity of disease. Am J Gastroenterol 96:2957–2961.

Wittmann DM, Krumsiek J, Saez-Rodriguez J, Lauffenburger DA, Klamt S, Theis FJ. 2009. Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling. BMC Syst Biol 3:98.

Yamada S, Shionoa S, Joob A, Yoshimura A. 2003. Control mechanism of JAK/STAT signal transduction pathway. FEBS Lett 534:190–196.

Yamamoto S, Matsumoto M, Tateno Y, Iinuma T, Matsumoto T. 1996. Quoit filter - a new filter based on mathematical morphology to extract the isolated shadow, and its application to automatic detection of lung cancer in X-ray CT. 13th International Conference on Pattern Recognition (ICPR'96), Vienna, Austria, 2:3-7.

Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED. 2004. Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics 20(18):3594–3603.

Zadeh LA. 1965. Fuzzy sets. Information and Control 8(3):338–353.

Zi ZK, Cho KH, Sung MH, Xia XF, Zheng JS, Sun ZR. 2005. In silico identification of the key components and steps in INF-γ induced JAK-STAT signaling pathway. FEBS Lett 579:1101–1108.

APPENDIX A

- Parameter values for oscillating transcription factor profile

$$U(s) = \frac{\omega_n^2}{s^2 + 2\varepsilon\omega_n s + \omega_n^2} \cdot \frac{T_\alpha}{s} e^{-\theta s}$$

$$A_1 = \frac{S_f S_n S_m C_{\mathrm{DNA}} T_a}{D_n D_m \Delta (D_n + S_f)}$$

$$A_2 = \frac{-S_n S_m C_{\mathrm{DNA}} \omega_n^2 T_a}{D_n \Delta (D_m - D_n)(D_n^2 - 2\varepsilon\omega_n D_n + \omega_n^2)}$$

$$A_3 = \frac{S_n S_m C_{\mathrm{DNA}} \omega_n^2 T_a}{\Delta (D_n + S_f)(D_m - D_n - S_f)((D_n + S_f)^2 - 2\varepsilon\omega_n (D_n + S_f) + \omega_n^2)}$$

$$A_4 = \frac{S_f S_n S_m C_{\mathrm{DNA}} \omega_n^2 T_a}{D_m \Delta (D_n - D_m)(D_m - D_n - S_f)(D_m^2 - 2\varepsilon\omega_n D_m + \omega_n^2)}$$

$$A_5 = \frac{\sqrt{A_7^2 + \left(\frac{A_6 - A_7\varepsilon\omega_n}{\omega_n\sqrt{1-\varepsilon^2}}\right)^2}}{\Delta}$$

$$A_6 = \frac{C_0(ad_1 + bd_0)}{bd_1^2 + bd_0^2}$$

$$A_7 = \frac{-C_0 d_1}{bd_1^2 + bd_0^2}$$

$$a = -\varepsilon\omega_n$$

$$b = \omega_n\sqrt{1-\varepsilon^2}$$

$$d_1 = -(a_3 + 4a)b^3 + (3a_3a^2 + 2a_2a + 4a^3 + a_1)b$$

$$d_0 = b^4 + a_3a^3 - (3a_3a + a_2 + 6a^2)b^2 + a_2a^2 + a^4 + a_1a$$

$$a_1 = (D_n^2 + D_n S_f) D_m$$

$$a_2 = D_n^2 + D_n S_f + 2D_n D_m + D_m S_f$$

$$a_3 = 2D_n + D_m + S_f$$

$$C_0 = S_f S_n S_m C_{\mathrm{DNA}} \omega_n^2 T_\alpha$$

$$\varphi = \arctan\frac{A_7\omega_n\sqrt{1-\varepsilon^2}}{A_6 - A_7\varepsilon\omega_n}$$

- Parameter values for monotone transcription factor profile

$$U(s) = \frac{1}{(\tau_1 s + 1)(\tau_2 s + 1)} \cdot \frac{T_\alpha}{s} e^{-\theta s}$$

$$B_1 = \frac{S_f S_n S_m C_{\mathrm{DNA}} T_a}{D_n D_m \Delta (D_n + S_f)}$$

$$B_2 = -\frac{S_n S_m C_{\mathrm{DNA}} T_a}{D_n \Delta (D_m - D_n)(-\tau_1 D_n + 1)(-\tau_2 D_n + 1)}$$

$$B_3 = \frac{S_n S_m C_{\mathrm{DNA}} T_a}{\Delta (D_n + S_f)(D_m - D_n - S_f)(-(D_n + S_f)\tau_1 + 1)(-(D_n + S_f)\tau_2 + 1)}$$

$$B_4 = -\frac{S_f S_n S_m C_{\mathrm{DNA}} T_a}{D_m \Delta (D_n - D_m)(-D_m \tau_1 + 1)(-D_m \tau_2 + 1)(D_n + S_f - D_m)}$$

$$B_5 = -\frac{S_f S_n S_m C_{\mathrm{DNA}} T_a \tau_1^4}{\Delta (D_n \tau_1 \text{-} 1)(D_m \tau_1 - 1)(\tau_1 - \tau_2)((D_n + S_f)\tau_1 - 1)}$$

$$B_6 = -\frac{S_f S_n S_m C_{\mathrm{DNA}} T_a \tau_2^4}{\Delta (D_n \tau_2 \text{-} 1)(\tau_2 - \tau_1)(D_m \tau_2 - 1)((D_n + S_f)\tau_2 - 1)}$$

- Parameter values for transcription factor profile with only one peak

$$U(s) = \frac{s - z_1}{s - p_1} \cdot \frac{1}{\tau_1 s + 1} \cdot \frac{T_\alpha}{s} e^{-\theta s}$$

$$C_1 = \frac{S_f S_n S_m C_{\mathrm{DNA}} z_1 T_a}{D_n D_m p_1 \Delta (D_n + S_f)}$$

$$C_2 = -\frac{S_n S_m C_{\mathrm{DNA}} T_a (D_n + z_1)}{D_n \Delta (D_m - D_n)(D_n + p_1)(1 \text{-} D_n \tau_1)}$$

$$C_3 = \frac{S_n S_m C_{\mathrm{DNA}} T_a (D_n + S_f + z_1)}{\Delta (D_n + S_f)(D_m - D_n - S_f)(D_n + S_f + p_1)(1 \text{-} (D_n + S_f)\tau_1)}$$

$$C_4 = -\frac{S_f S_n S_m C_{\mathrm{DNA}} T_a (D_m + z_1)}{D_m \Delta (D_n - D_m)(D_m + p_1)(D_n + S_f - D_m)(1 \text{-} D_{\mathrm{m}} \tau_1)}$$

$$C_5 = \frac{S_f S_n S_m C_{\mathrm{DNA}} T_a (p_1 - z_1)}{p_1 \Delta (p_1 + D_n)(p_1 + D_m)(1 + p_1 \tau_1)(p_1 + D_n + S_f)}$$

$$C_6 = -\frac{S_f S_n S_m C_{\mathrm{DNA}} T_a \tau_1^3 (z_1\tau_1 + 1)}{\Delta(1 + p_1\tau_1)(D_n\tau_1 - 1)(D_m\tau_1 - 1)((D_n + S_f)\tau_1 - 1)}$$

APPENDIX B

- Model equations

$$\dot{x}_1 = -k_{1p}x_1u + k_{2p}x_2 + k_{17p}x_{26} + k_{11p}x_{10}$$
$$\dot{x}_2 = k_{1p}x_1u - k_{2p}x_2 - k_{3p}x_2x_3 + k_{4p}x_4$$
$$\dot{x}_3 = -k_{3p}x_2x_3 + k_{4p}x_4 + k_{11p}x_{10} + k_{20p}x_{29}$$
$$\dot{x}_4 = k_{3p}x_2x_3 - k_{4p}x_4 - k_{5p}x_4x_5 + k_{6p}x_6$$
$$\dot{x}_5 = -k_{5p}x_4x_5 + k_{6p}x_6 + k_{11p}x_{10} + k_{20p}x_{29}$$
$$\dot{x}_6 = k_{5p}x_4x_5 - k_{6p}x_6 - k_{7p}x_6x_7 + k_{8p}x_8$$
$$\dot{x}_7 = -k_{7p}x_6x_7 + k_{8p}x_8 + k_{11p}x_{10} + k_{20p}x_{29}$$
$$\dot{x}_8 = k_{7p}x_6x_7 - k_{8p}x_8 - k_{9p}x_8x_9 + k_{10p}x_{10} - k_{15p}x_8x_{25} + k_{16p}x_{26}$$
$$\dot{x}_9 = -k_{9p}x_8x_9 + k_{10p}x_{10} + k_{14p}x_{14}$$
$$\dot{x}_{10} = k_{9p}x_8x_9 - k_{10p}x_{10} - k_{11p}x_{10}$$
$$\dot{x}_{11} = k_{11p}x_{10} - k_3x_{11} - T_R k_2x_{11}x_{17} - k_{deg}x_{11} - a_2x_{11}x_{19} + t_1x_{13} - a_3x_{11}x_{22} + t_2x_{14}$$
$$\dot{x}_{12} = k_3x_{11} + T_R k_2x_{11}x_{17} - k_{deg}x_{12}$$
$$\dot{x}_{13} = a_2x_{11}x_{19} - t_1x_{13}$$
$$\dot{x}_{14} = a_3x_{11}x_{22} - t_2x_{14}$$
$$\dot{x}_{15} = c_{6a}x_{22} - a_1x_{15}x_{19} + t_2x_{14} - i_1x_{15}$$
$$\dot{x}_{16} = i_1k_vx_{15} - a_1x_{20}x_{16}$$
$$\dot{x}_{17} = c_4x_{18} - c_5x_{17}$$
$$\dot{x}_{18} = c_2 + c_1x_{16} - c_3x_{18}$$
$$\dot{x}_{19} = -a_2x_{11}x_{19} - a_1x_{19}x_{15} + c_{4a}x_{21} - c_{5a}x_{19} - i_{1a}x_{19} + e_{1a}x_{20}$$
$$\dot{x}_{20} = -a_1x_{16}x_{20} + i_{1a}k_vx_{19} - e_{1a}k_vx_{20}$$
$$\dot{x}_{21} = c_{2a} + c_{1a}x_{16} - c_{3a}x_{21}$$
$$\dot{x}_{22} = a_1x_{19}x_{15} - c_{6a}x_{22} - a_3x_{11}x_{22} + e_{2a}x_{23}$$
$$\dot{x}_{23} = a_1x_{20}x_{16} - e_{2a}k_vx_{23}$$
$$\dot{x}_{24} = c_{2c} + c_{1c}x_{16} - c_{3c}x_{24} - k_{28p}x_{24}x_{33}$$
$$\dot{x}_{25} = -k_{15p}x_8x_{25} + k_{16p}x_{26} + k_{20p}x_{29}$$
$$\dot{x}_{26} = k_{15p}x_8x_{25} - k_{16p}x_{26} - k_{17p}x_{26}$$
$$\dot{x}_{27} = k_{17p}x_{26} - k_{18p}x_{27}x_{28} + k_{19p}x_{29}$$

$$\dot{x}_{28} = -k_{18p}x_{27}x_{28} + k_{19p}x_{29}$$

$$\dot{x}_{29} = k_{18p}x_{27}x_{28} - k_{19p}x_{29} - k_{20p}x_{29}$$

$$\dot{x}_{30} = k_{20p}x_{29} - k_{21p}x_{30}x_{31} + k_{22p}x_{32} + k_{23p}x_{32}$$

$$\dot{x}_{31} = -k_{21p}x_{30}x_{31} + k_{22p}x_{32}$$

$$\dot{x}_{32} = k_{21p}x_{30}x_{31} - k_{22p}x_{32} - k_{23p}x_{32}$$

- State variables of the model and their initial values:

Name	Species	Initial values (μM)
x_1	TNFR1	0.1
x_2	TNF-α/TNFR	0
x_3	TRADD	0.15
x_4	TNF-α/TNFR1/TRADD	0
x_5	TRAF2	0.1
x_6	TNF-α/TNFR1/TRADD/TRAF2	0
x_7	RIP-1	0.1
x_8	TNF-α/TNFR1/TRADD/TRAF2/RIP-1	0
x_9	IKKn	0.2
x_{10}	TNF-α/TNFR1/TRADD/TRAF2/RIP-1/IKKn	0
x_{11}	IKKa	0
x_{12}	inactive IKK	0
x_{13}	cytoplasmic IKK|IkBa complex	0
x_{14}	cytoplasmic IKK|IkBa|NF-κB complex	0
x_{15}	free cxtoplasmic NF-κB	0.0003
x_{16}	free nuclear NF-κB	0.0023
x_{17}	cytoplasmic A20	0.0048
x_{18}	A20 transcription	0
x_{19}	free cytoplasmic IkBa	0.0025
x_{20}	free nuclear IkBan	0.0034
x_{21}	IkB transcription	0
x_{22}	cytoplasmic IkBa|NF-κB complex	0.0592
x_{23}	Nuclear IkBa|NF-κB complex	0.0001
x_{24}	Control gene mRNA level or c-IAP	0
x_{25}	FADD	0.1
x_{26}	TNF-α/TNFR1/TRADD/TRAF2/RIP-1/FADD	0
x_{27}	TRADD/TRAF2/RIP-1/FADD	0
x_{28}	Caspase-8	0.08
x_{29}	TRADD/TRAF2/RIP-1/FADD/caspase-8	0
x_{30}	Caspase-8*	0
x_{31}	Caspase-3	0.2
x_{32}	Caspase-8*/caspase-3	0

Name	Species	Initial values (μM)
x_{33}	Caspase-3*	0
x_{34}	DNA-fragmentation	0
x_{35}	Caspase-3*/c-IAP	0
x_{36}	DNA intact	0.8
x_{37}	Caspase-3*/DNA	0

note: u is the concentration of TNF-α, ng/ml. The molecule weight of TNF-α is 17 kDa. The unit ng/ml can be converted to μM by dividing by 17×10^3. y is the system output NF-κB after being scaled by k_r in units of μM.

- Values of the parameters

Name	Value	Name	Value
k_v	5	k_{1p}	0.0740 (0.185)
AB^*	1	k_{15p}	0.185
c_1	$5\times10^{-7}\ AB$	k_{2p}	0.00125
c_2	0	k_{16p}	0.00125
c_3	0.0104 (0.0004)	k_{3p}	0.185
c_4	0.5	k_{17p}	0.37
c_5	0.0003	k_{4p}	0.00125
k_1	0.0025	k_{18p}	0.5
k_2	0.1	k_{5p}	0.185
k_3	0.0015	k_{19p}	0.2
k_{deg}	0.000125	k_{6p}	0.00125
a_2	0.2	k_{20p}	0.1
a_1	0.5	k_{7p}	0.185
a_3	1.	k_{21p}	0.1
t_1	0.1	k_{8p}	0.00125
t_2	0.1	k_{22p}	0.06
AA^*	1	k_{9p}	0.185
c_{1a}	$5\times10^{-7}\ AA$	k_{23p}	100
c_{2a}	0	k_{10p}	0.00125
c_{3a}	0.0004	k_{24p}	0.185
c_{4a}	0.5	k_{11p}	0.37
c_{5a}	0.0001	k_{25p}	0.00125
c_{6a}	0.00002	k_{12p}	0.014
i_1	0.0025	k_{26p}	0.37
e_{2a}	0.01	k_{13p}	0.00125

Name	Value	Name	Value
i_{1a}	0.001	k_{14p}	0.37
e_{1a}	0.0005	k_{28p}	0.5
c_{1c}	5×10^{-7}	p	1.75
c_{2c}	0	T_r^*	1
c_{3c}	0.0004	k_r	2.5

* Note: 1) AA = 1 refers to wt cell, while AA = 0 refers to IkBa deficient cell
2) AB = 1 refers to wt cell, while AB = 0 refers to A20 deficient cell
3) T_r = 0 when TNF-α is off, while T_r = 1 when TNF-α is on
4) Values in brackets refer to the model fit to the experimental data

APPENDIX C

- Model equations:

$$\dot{x}_1 = p_1 u^2 R^2 - p_2 x_1 - p_3 x_1 x_2 + p_4 x_3 + p_5 x_3 - p_{11} x_1 x_5 + p_{12} x_6 - p_{14} x_1 x_7 + p_{15} x_8$$
$$\dot{x}_2 = -p_3 x_1 x_2 + p_4 x_3 + 2 p_6 x_4$$
$$\dot{x}_3 = p_3 x_1 x_2 - p_4 x_3 - p_5 x_3$$
$$\dot{x}_4 = p_5 x_3 / 2 - p_6 x_4$$
$$\dot{x}_5 = p_7 x_4 \mathrm{Step}(t - p_8) / (p_9 + x_4) - p_{10} x_5 - p_{11} x_1 x_5 + p_{12} x_6 + p_{13} x_6$$
$$\dot{x}_6 = p_{11} x_1 x_5 - p_{12} x_6 - p_{13} x_6$$
$$\dot{x}_7 = -p_{14} x_1 x_7 + p_{15} x_8$$
$$\dot{x}_8 = p_{14} x_1 x_7 - p_{15} x_8$$
$$\dot{x}_9 = p_{16} x_8 x_{10} / (p_{17} + x_{10}) - p_{18} x_9$$
$$\dot{x}_{10} = -p_{16} x_8 x_{10} / (p_{17} + x_{10}) + p_{18} x_9$$
$$\dot{x}_{11} = p_{13} x_6$$
$$\dot{x}_{12} = -2\, p_{19}\, x_9 x_{12}^2$$
$$\dot{x}_{13} = p_{19}\, x_9 x_{12}^2$$

- State variables of the model and their initial values:

Name	Component	Initial value (nM)
x_1	$(\text{IL6-gp80-gp130-JAK})^*_2$	0
x_2	STAT3C	1000
x_3	$(\text{IL6-gp80-gp130-JAK})^*_2$-STAT3C	0
x_4	STAT3N^*- STAT3N^*	0
x_5	SOCS3	0
x_6	$(\text{IL6-gp80-gp130-JAK})^*_2$-SOCS3	0
x_7	SHP2	100
x_8	$(\text{IL6-gp80-gp130-JAK})^*_2$-SHP2-sum	0
x_9	Erk-PP	0
x_{10}	Erk	16468
x_{11}	$(\text{IL6-gp80-gp130-JAK})_2$	0
x_{12}	C/EBPβi	40.493
x_{13}	C/EBPβn	0
u	IL-6	3.83 (i.e.,100 ng/ml)
R	Receptor	4

- Values of the parameters

Name	Physical Interpretation	Value
p_1	Forward rate constant for Reaction #1	2.336e-005
p_2	Backward rate constant for Reaction #1	0.002
p_3	Forward rate constant for Reaction #2	0.0138
p_4	Backward rate constant for Reaction #2	1.502
p_5	Forward rate constant for Reaction #3	0.273
p_6	Forward rate constant for Reaction #4	3.282e-004
p_7	Maximum rate for Reaction #5	0.023
p_8	Time delay for Reaction #5	1290
p_9	Michaelis-Menten constant for Reaction #5	50.6
p_{10}	Forward rate constant for Reaction #6	2.067e-004
p_{11}	Forward rate constant for Reaction #7	16.52
p_{12}	Backward rate constant for Reaction #7	0.0400
p_{13}	Forward rate constant for Reaction #8	0.0023
p_{14}	Forward rate constant for Reaction #9	4.059e-004
p_{15}	Backward rate constant for Reaction #9	5.086e-004
p_{16}	Maximum rate for Reaction #10	16.00
p_{17}	Michaelis-Menten constant for Reaction #10	5.115e+003
p_{18}	Forward rate constant for Reaction #11	1.198e-005
p_{19}	Forward rate constant for Reaction #12	1.0 e-006

* Note: the reaction numbers, *#n*, are consistent with those shown in Fig. 10. First order rate constants have units of 1/s and second order rate constants of [nM^{-1} s^{-1}].

VITA

Zuyi Huang received his Bachelor of Engineering degree in thermal engineering from Tsinghua University, Beijing, China, in July 2001. He received his Master of Science degree in thermal engineering from the same university in July 2004. He entered the graduate program in chemical engineering at Texas A&M University in May 2006. He received his Ph.D. degree from Texas A&M University in August 2010. His research interests include mathematical modeling of signal transduction pathways, statistical multivariate data analysis, model reduction of nonlinear models, and fuzzy modeling.

Dr. Huang can be reached c/o Dr. Juergen Hahn at Artie McFerrin Department of Chemical Engineering, Texas A&M University, 3122 TAMU, College Station, TX 77843. His email address is huangzuyi@gmail.com.