Regularization of Inverse Problems to Determine Transcription Factor Profiles from Fluorescent Reporter Systems

Loveleena Bansal, Yunfei Chu, Carl Laird, and Juergen Hahn
Artie McFerrin Dept. of Chemical Engineering, Texas A&M University, College Station, TX 77843

DOI 10.1002/aic.13782
Published online March 15, 2012 in Wiley Online Library (wileyonlinelibrary.com).

Signal transduction pathways are characterized by complex biochemical reactions which involve a large number of proteins. The availability and quality of experimental data pose challenges for identifying the role of individual proteins in these pathways. To address this issue, this article formulates and solves an inverse problem to determine the dynamics of transcription factors from fluorescence intensity measurements of green fluorescent protein (GFP) reporter systems. In the presented approach, a model describing transcription and translation of GFP is discretized and concentrations of transcription factor are estimated at discrete time points. Unlike previous studies, this approach has no restrictions with regard to a particular shape of the profiles. However, the resulting inverse problem is ill-conditioned and requires the use of regularization techniques. Two regularization methods—truncated singular value decomposition and Tikhonov regularization—are investigated in this work and the characteristics of the results obtained are discussed in detail. © 2012 American Institute of Chemical Engineers *AIChE J,* 58: 3751–3762, 2012
Keywords: inverse problem, truncated singular value decomposition, Tikhonov regularization, transcription factor dynamics, green fluorescent protein

Introduction

Transcription factors (TF) are key elements of signal transduction pathways as they are involved in initiation of the transcription/translation process leading to the formation of new proteins in the cell. Thus, a quantitative description of transcription factor dynamics can aid in understanding the response of cells to external stimuli.[1,2] The activation of TF has been conventionally monitored using protein binding methods like western blot analysis or chromatin immunoprecipitation. However, these techniques provide only qualitative or semiquantitative data and are destructive measurement techniques, i.e., the same sample cannot be monitored continuously over time.

A number of researchers in the last two decades have used fluorescence-based reporter systems for continuous and noninvasive monitoring of gene expression and transcriptional activity.[3,4] Using these techniques, the underlying dynamics of TF cannot be directly monitored but the fluorescence of proteins, such as the green fluorescent protein (GFP), observed from fluorescence microscopy or a fluorescent plate reader, can be used as an indicator of the activation of TF. However, the relationship between the concentration of the TF and the observed fluorescence is not straightforward as it involves dynamic processes dealing with transcription, translation, and post-translational modification of GFP.[4] A number of mathematical models describing these processes have been developed,[5–9] and these models have been used for estimating the mRNA or transcription dynamics from gene expression data. In one of the studies,[10] the concentrations of the SOS transcriptional repressor for the SOS DNA repair system in *E. coli* were estimated but the post-translation modifications of GFP were not explicitly taken into account in the dynamic model. Other works assumed a certain nature of the dynamic profile of a compound and then estimated the parameters to characterize the profile.[5,6,11] One drawback of this approach is that it restricts the functional form of the estimated profiles. This type of approach was later extended to several different functional forms,[12] however, a general approach for computing the transcription dynamics has not been presented so far. Thus, it is the main aim of this article to develop an inverse problem formulation such that dynamic transcription factor profiles of any shape can be estimated from the fluorescent intensity profiles of fluorescent reporter systems.

A previously developed ordinary differential equation (ODE) model[11] which describes transcription, translation, and fluorophore formation of GFP has been used in this work. The input to this ODE model is the time-dependent concentration of a transcription factor and the observed fluorescence is treated as the output. Although this inverse problem appears to be relatively simple at first glance as it only involves one dynamic input, one dynamic output, and three nonlinear differential equations, there are significant challenges that arise from the time scales of the input/output and the time constants of the model as well as the noise level of the measurements. The resulting discrete inverse problem is also highly ill-conditioned. Because of this, it is one important component of this work to investigate regularization schemes

Additional Supporting Information may be found in the online version of this article.

Correspondence concerning this article should be addressed to J. Hahn at hahn@tamu.edu.

which can deal with this ill-conditioned inverse problem and also filter the effect of noisy measurements on the estimated input profile. Furthermore, challenges arise since experimental data might be missing or only available at large time intervals when compared with the transcription factor dynamics. Both these issues will also be addressed in this work.

Regularization has been widely used for solving ill-conditioned inverse problems in a variety of areas including, but not limited to, electrocardiography,[13] geophysics,[14] and electrical impedance tomography.[15] A number of regularization methods for solving discrete inverse problems such as truncated singular value decomposition (TSVD),[16] Tikhonov regularization,[17,18] total least squares,[19] and several iterative methods[20,21] exist. Among these techniques, there is no method that performs best for all types of inverse problems. Because of this, the two most commonly used methods TSVD and Tikhonov regularization—have been used in this work for the solution of the presented inverse problem. These regularization techniques have also been implemented with non-negative constraints to obtain transcription factor concentrations which are biologically possible. This inverse problem has not been previously investigated in a discrete regularized form, thus it is not known that which regularization method is more suited for its solution. Thus, a comparison of the results obtained by the two methods is made in this work. The reason for focusing on only these two regularization techniques is that total least squares is intended for cases where the coefficient matrix contains large perturbations[22,23] while iterative methods cater to large scale problems.[20] However, these two situations do not arise for this inverse problem dealing with computation of the transcription factor profiles from fluorescent intensity profiles discussed in this article.

This work includes both the theoretical and the practical aspects of including regularization for the solution of the discrete inverse problem for computing transcription factor profiles. The next section discusses background material associated with the ODE model describing the transcription/translation process and a brief overview of regularization methods is done. Then, the ODE model is recast as a linear regression model and the theoretical formulation for applying regularization methods for the solution of the inverse problem is discussed. The presented technique is illustrated using two case studies. In the first case study, simulated data is used to compare the results obtained for the two regularization methods. Then the technique is applied to experimental data for estimating the profiles of the transcription factor STAT3. These data are available in the form of fluorescence microscopy images obtained from the continuous stimulation of hepatocytes with 100 ng/ml of IL-6.

Preliminaries

Model describing transcription, translation, and activation of GFP

The overall scheme for activation of TF and the observed fluorescence intensity in fluorescent protein reporter systems is shown in Figure 1. Stimulation of signaling pathways by cytokines leads to the activation of TF which translocates to the nucleus of the cell. The TF bind to the DNA in the nucleus and initiate the transcription/translation process which leads to fluorescence via formation of fluorescent proteins. Estimation of transcription factor dynamics from the fluorescence intensity profiles requires a model describing this transcription/translation process.

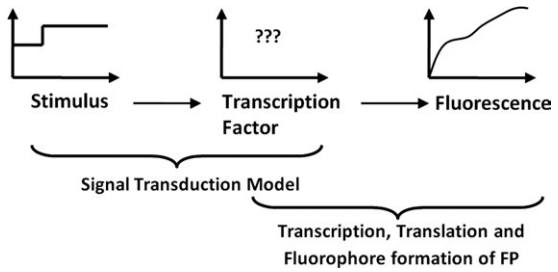

Figure 1. Dynamic scheme of signaling pathway and gene expression process.

An ODE model describing transcription, translation, and activation of GFP[9] is used in this work. This model consists of three ODEs which result from the component balances of the amounts of m-RNA, the nonfluorescent form of GFP, and the fluorescent form of GFP. The model had been modified[11] to take in account of the constant reporter DNA levels due to stable integration of the reporter plasmid into the genomic DNA and the effect of transcription factor concentrations on the transcription rate. The resulting model is given by

$$\begin{aligned}\frac{dm}{dt} &= S_m p \frac{C_{TF}}{c + C_{TF}} - D_m\, m \\ \frac{dn}{dt} &= S_n m - D_n n - S_f\, n \\ \frac{df}{dt} &= S_f\, n - D_n f\end{aligned} \tag{1}$$

where C_{TF} is the concentration of the transcription factor in the nucleus, m is the mRNA concentration, n is the concentration of nonfluorescent GFP, and f is the concentration of fluorescent GFP. The parameters and their constant values are S_m is the transcription rate which is constant for a transcription factor and has a value of 373 h^{-1} for NF-kB and has been re-estimated[24] for STAT3 and C/EBP-β to be 548 h^{-1} and 329.35 h^{-1}, respectively; p is the amount of DNA with a value of 5 nM; c has a value of 108 nM; D_m is the constant mRNA degradation rate that equals 0.45 h^{-1}; S_n is the translation rate and is equal to 780 h^{-1}; D_n is the protein degradation rate which equals 0.5 h^{-1}; S_f is the fluorophore formation rate which depends on the GFP variant used and it has a value of 0.347 h^{-1} in this study.

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

$$I = f/\Delta \tag{2}$$

where Δ has a value of 2.5562×10^4 nM. The initial conditions for this system are $m(0) = 0$ nM, $n(0) = 0$ nM, and $f(0) = 0$ nM. Though, the mRNA levels of the fluorescence proteins may be nonzero initially but the concentrations are very low[6] and thus they can safely be assumed to be zero. It is the aim of this work to calculate the transcription factor profile C_{TF} from the measured fluorescent intensity I.

Review of regularization methods

A main challenge in solving discrete inverse problems is that the problem can be ill-conditioned such that small perturbations in the measurements can produce large variations in the solution.[21,25] Regularization procedures need to be included in the regression formulation to ensure stable parameter

estimates are obtained. In this regard, two commonly used regularization methods, i.e., TSVD and Tikhonov regularization, are reviewed in this subsection as they will be used for solution of the inverse problem presented in this article.

Assume that a linear regression model is given by

$$\tilde{\mathbf{y}} = \mathbf{H}\mathbf{u} + \varepsilon \tag{3}$$

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

The solution for the unknown $\mathbf{u}$ in Eq. 3 can be computed by

$$\hat{\mathbf{u}} = \mathbf{H}^{+}\tilde{\mathbf{y}} \tag{4}$$

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

$$\begin{aligned} \mathbf{H}^{+} &= (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}, \qquad p \geq q \\ \mathbf{H}^{+} &= \mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{H}^{\mathrm{T}})^{-1}, \qquad p < q \end{aligned} \tag{5}$$

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

Truncated Singular Value Decomposition. The solution shown in Eq. 4 can be represented in an alternate form by calculating the singular value decomposition (SVD) of the transfer matrix, $\mathbf{H} = \mathbf{U}\mathbf{W}\mathbf{V}^{\mathrm{T}}$ where $\mathbf{U} \in R^{p \times p}$, $\mathbf{V} \in R^{q \times q}$ and $\mathbf{W} \in R^{p \times q}$. These matrices satisfy

$$\begin{aligned} &\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathrm{I}_p \\ &\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathrm{I}_q \\ &\mathbf{W}_{ij} = \begin{cases} w_i & \forall\, i = j \\ 0 & \forall\, i \neq j \end{cases} \end{aligned} \tag{6}$$

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

$$\mathbf{H}^{+} = \mathbf{V}\mathbf{W}^{+}\mathbf{U}^{\mathrm{T}} \tag{7}$$

where $\mathbf{W}^{+}$ is the pseudo inverse of $\mathbf{W}$, which can be evaluated by doing the reciprocal of all nonzero diagonal elements and transposing the matrix. Substituting Eq. 7 in 4, the solution can be written in the following decomposed spectral form

$$\hat{\mathbf{u}} = \sum_{i=1}^{r} \frac{\mathbf{u}_i^T \tilde{\mathbf{y}}}{w_i}\mathbf{v}_i \tag{8}$$

where $\mathbf{u}_i \in R^p$ and $\mathbf{v}_i \in R^q$ are columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. The components corresponding to small singular values in Eq. 8 are responsible for large errors in the solution of discrete linear inverse problems.[21] Using TSVD regularization, the solution is obtained by considering only the first k components of the singular value decomposition corresponding to large singular values

$$\hat{\mathbf{u}}_w = \sum_{i=1}^{k} \frac{\mathbf{u}_i^{\mathrm{T}} \tilde{\mathbf{y}}}{w_i}\mathbf{v}_i \tag{9}$$

The choice of the regularization parameter k can be made on the basis of the discrete Picard condition.[16] According to this condition, the numerator $\mathbf{u}_i^{\mathrm{T}}\tilde{\mathbf{y}}$ should decay faster than the singular values w_i such that the overall norm of the SVD components $|\mathbf{u}_i^{\mathrm{T}}\tilde{y}/w_i|$ is small. For practical applications, $|\mathbf{u}_i^{\mathrm{T}}\tilde{\mathbf{y}}|$ and the singular values are plotted for all the SVD components of the sum given in Eq. 8 and truncation parameter is chosen until the Picard condition is satisfied. Also, to obtain non-negative solutions for quantities that cannot be negative, additional non-negativity constraints need to be applied in the regularization formulation. There have been few instances when truncated SVD has been implemented with non-negative constraints. These formulations range from simply setting the negative values in the estimated solutions to zero[26] to mathematically more rigorous formulations involving quadratic programming problem with bounds.[27,28]

Tikhonov Regularization. A least squares formulation seeks to minimize the norm of the residual between the estimated and the measured values given by $\|\tilde{\mathbf{y}} - \mathbf{H}\mathbf{u}\|_2^2$. Tikhonov regularization, also known as Ridge regression in statistics,[29] adds a regularization term to this residual, which results in the following formulation

$$\min_{\mathbf{u}} \|\tilde{\mathbf{y}} - \mathbf{H}\mathbf{u}\|_2^2 + \lambda\|\mathbf{L}\mathbf{u}\|_2^2 \tag{10}$$

Here, λ is a regularization parameter which denotes the weight of the regularization term and $\mathbf{L}\mathbf{u}$ is a finite difference approximation that is proportional to the derivatives of u.[18] The term $\|\mathbf{L}\mathbf{u}\|_2^2$ tries to minimize the effect of noise components by minimizing the norm of the solution. This term also aids in the solution of under-determined system of algebraic equations by decreasing the degrees of freedom. The explicit solution for this minimization problem is given by

$$\hat{\mathbf{u}}_\lambda = (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{L}^T\mathbf{L})^{-1}(\mathbf{H}^T\tilde{\mathbf{y}}) \tag{11}$$

The regularization parameter can be chosen with the help of the L-curve[17] which is a plot of the norm of the residual vs. the regularization term for various values of the parameter. The two norms vary monotonically with the regularization parameter with opposite trends and result in a L-shaped curve. The parameter is chosen around the corner of this L-curve to maintain a balance between the residual and the norm of the solution. This rule of thumb results from the fact that little is gained in terms of minimizing the norm of the solution by increasing the parameter λ from the one at the corner value, or with respect to minimizing the residual by decreasing λ significantly below the corner value due to the characteristic L-shape of the curve.

Tikhonov regularization has commonly been applied with non-negative constraints.[30,31] This involves adding the following bounds to the formulation shown in Eq. 10

$$\mathbf{u} \geq \mathbf{0} \tag{12}$$

This results in a quadratic programming problem that can be solved with the aid of optimization solvers that are now commercially available.

Procedure for Solving the Inverse Problem to Obtain Transcription Factor Profiles

This section formulates the inverse problem for estimating transcription factor profiles from fluorescence intensity profiles. The continuous time ODE model describing

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

Inverse problem formulation

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

$$\begin{aligned} &\min_{\hat{C}_{\mathrm{TF}}(t)} \sum_{i=0}^{m} (\tilde{y}_i - y_i)^2 \\ &\textbf{\textit{s.t.}} \quad y_i = g(\hat{C}_{\mathrm{TF}}, T_i) \;\; \forall \;\; T_i = \{T_0, T_1, \ldots, T_m\} \\ &\qquad t \in [t_0, t_n] \end{aligned} \tag{13}$$

where $\hat{C_{\mathrm{TF}}}(t)$ is the continuous transcription factor profile for $t \in [t_0,t_n]$, $\tilde{y_i}$ is the discrete fluorescence intensity measurement at time T_i, y_i is the estimated intensity at time T_i using $\hat{C_{\mathrm{TF}}}(t)$ and the model describing the transcription/translation process given by Eqs. 1 and 2. This dynamic model is denoted by $y_i = g(\hat{C_{\mathrm{TF}}},T_i)$. The range of the time interval in which measurements are available is $[T_0,T_m]$ and consists of $m + 1$ sampling points. The sampling step size for these measurements need not be uniform in this formulation.

The optimization problem (13) is nontrivial to solve as the equality constraint consists of 3 ODEs (see Eq. 1). This continuous formulation would result in an infinite dimensional inverse problem. To avoid this, if a functional form is assumed for the profile of $\hat{C_{\mathrm{TF}}}(t)$, it would restrict the shape of the estimated profiles. Thus, a different approach is used in this work. It is assumed that the profile for $\hat{C_{\mathrm{TF}}}(t)$ is piecewise constant over a discretization interval and only changes between the discretization points. Furthermore, the ODE model representing the first constraint is discretized resulting in algebraic equations describing the model. Thus, the transcription factor profile is discrete and its values are estimated at each discrete time point. It should be noted that the discretization of the transcription factor profile does not have to be the same as the time points at which measurements are available.

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

$$u(t) = \frac{C_{\mathrm{TF}}(t)}{c + C_{\mathrm{TF}}(t)} \tag{14}$$

eliminates the nonlinearity in the model given by Eqs. 1 and 2 and results in the linear dynamic system

$$\begin{aligned} \frac{dm}{dt} &= S_m p u - D_m m \\ \frac{dn}{dt} &= S_n m - D_n n - S_f n \\ \frac{df}{dt} &= S_f n - D_n f \end{aligned} \tag{15}$$

$$y = f/\Delta \tag{16}$$

where y is substituted for the output fluorescent intensity I. This system can be represented in the state-space form as

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t) \\ y(t) &= \mathbf{C}\mathbf{x}(t) + \varepsilon(t) \end{aligned} \tag{17}$$

where the measurement noise is denoted by $\varepsilon(t)$ and state vector and system matrices are given by

$$\mathbf{x} = [m \quad n \quad f]^{\mathrm{T}} \tag{18}$$

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} -D_m & 0 & 0 \\ S_n & -D_n - S_f & 0 \\ 0 & S_n & -D_n \end{bmatrix} \\ \mathbf{B} &= [S_m p \quad 0 \quad 0]^{\mathrm{T}} \\ \mathbf{C} &= [0 \quad 0 \quad 1/\Delta] \end{aligned} \tag{19}$$

As the system given by Eq. 15 is a linear dynamic system, it has a closed-form solution given by

$$y(T) = \mathbf{C} \int_0^T e^{\mathbf{A}(T-\tau)} u(\tau) d\tau \, \mathbf{B} \tag{20}$$

which can be used for discretizing the model.

The input $u(t)$ to the system is discretized and is assumed to be constant between two consecutive time points where discretization was performed according to the scheme shown in Figure 2. If the input and the output are sampled at different times as shown in Figure 2, then the discrete output is

$$\begin{aligned} y(T_i) &= \mathbf{C} \left(\sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} e^{\mathbf{A}(T_i-\tau)} u_{j-1} d\tau \right. \\ &\quad \left. + \int_{t_k}^{T_i} e^{\mathbf{A}(T_i-\tau)} u_k \, d\tau \right) \mathbf{B} \qquad \text{for} \quad t_k < T_i \le t_{k+1} \\ &= \sum_{j=1}^{k} \mathbf{C} \left(\int_{t_{j-1}}^{t_j} e^{\mathbf{A}(T_i-\tau)} d\tau \right) \mathbf{B} \, u_{j-1} + \mathbf{C} \left(\int_{t_k}^{T_i} e^{\mathbf{A}(T_i-\tau)} d\tau \right) \mathbf{B} \, u_k \end{aligned} \tag{21}$$

This solution can be represented in the form of a linear regression model as described in Eq. 3, in which the transfer matrix $\mathbf{H}$, the output vector $\mathbf{y}$, the input vector $\mathbf{u}$, and the noise vector ε are given by

$$\mathbf{H}_{ij} = \begin{cases} \mathbf{C}(\int_{t_{j-1}}^{t_j} e^{\mathbf{A}(T_i-\tau)} d\tau)\mathbf{B} & t_j < T_i \\ \mathbf{C}(\int_{t_j}^{T_i} e^{\mathbf{A}(T_i-\tau)} d\tau)\mathbf{B} & t_{j-1} < T_i \le t_j \\ 0 & T_i \le t_{j-1} \end{cases} \tag{22}$$

$$\begin{aligned} \mathbf{y} &= [y_0 \quad y_1 \quad \cdots \quad y_m]^T \\ \mathbf{u} &= [u_0 \quad u_1 \quad \cdots \quad u_{n-1}]^T \\ \varepsilon &= [\varepsilon_0 \quad \varepsilon_1 \quad \cdots \quad \varepsilon_m]^T \end{aligned} \tag{23}$$

The above formulation does not require the measurements to be available in the same time interval as the input. However, if the measurements are available in the interval $[T_0,T_m]$, the input can be calculated for the interval $[t_0,t_n]$, such that $t_n \le T_m$. In this formulation, the discrete values of the output fluorescence intensity $\mathbf{y}$ are directly related to the input $\mathbf{u}$ which is a function of the transcription factor concentration (see Eq. 14). The $\mathbf{H} \in R^{(m+1) \times n}$ matrix usually has more columns than rows, i.e., $n > m + 1$, because experimental data are available only at a few time points but the input profile

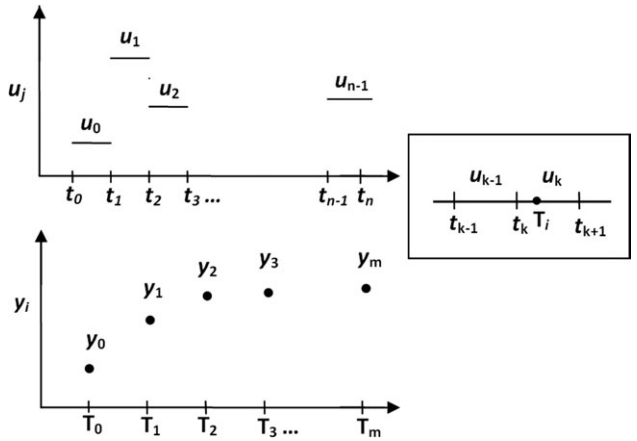

Figure 2. Discretization of the ODE model with zero-order hold for the input.

needs to be estimated at several points in time. If any data values are missing, the corresponding row can be removed from the transfer matrix and the input vector remains unchanged. The method for evaluating the integrals shown in Eq. 22 is given in the Appendix.

The optimization problem from Eq. 13 can now be formulated as

$$\min_{\{c_{TFj}\}_{j=0}^{j=n-1}} \sum_{i=0}^{m} (\tilde{y}_i - y_i)^2$$
$$s.t. \quad y_i = \mathbf{H}_i \mathbf{u} \qquad \forall \; i = \{0, 1 \ldots, m\}$$
$$u_j = \frac{C_{TFj}}{c + C_{TFj}} \; \forall \, j = \{0, 1, \ldots, n-1\} \quad (24)$$

where $y_i = y(T_i)$, u_j and C_{TF_j} are the constant values of the input and the transcription factor in the discrete interval $t_j \leq t \leq t_{j+1}$ and $\mathbf{H}_i \in R^n$ is a row of the transfer matrix evaluated in Eq. 22. This formulation includes algebraic equality constraints instead of the system of ODEs which had to be solved for the original formulation in Eq. 13. This inverse problem is found to be highly ill-conditioned. Because of this, there is a need to include a regularization procedure for the solution of this inverse problem.

Application of regularization methods and solution of the inverse problem

This section explains how regularization methods have been integrated into the solution of the inverse problem for computing transcription factor profiles from fluorescence intensity profiles and the inverse solution has been calculated.

Discretization of the continuous ODE model results in an under-determined linear regression model given by Eqs. 22 and 23. This under-determined system forms a part of the optimization problem described by Eq. 24. This optimization problem has been transformed to solve for the unknown input $\mathbf{u}$ instead of C_{TF} such that the following formulation is obtained

$$\min_{\mathbf{u}} \|\tilde{\mathbf{y}} - \mathbf{y}\|_2^2$$
$$\text{s.t.} \quad \mathbf{y} = \mathbf{H}\mathbf{u} \quad (25)$$

where $\tilde{\mathbf{y}} = [\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_m]^T$, $\mathbf{y} = [y_1, y_2, \ldots, y_m]^T$ and $\mathbf{u}$ is given in Eq. 23. Transcription factor profiles are then obtained from the estimated input using the transformation from Eq. 14

$$C_{TF}(t_j) = c \frac{u(t_j)}{1 - u(t_j)} \quad \forall \; t_0 \leq t_j \leq t_{n-1} \quad (26)$$

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

$$\bar{\mathbf{u}} = \mathbf{L}\,\mathbf{u} \quad (27)$$

where $\mathbf{L}$ is a finite element approximation matrix of the first order derivative

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 1 & 0 \\ & & \cdots & & \\ 0 & 0 & & -1 & 1 \end{bmatrix}_{n \times n} \quad (28)$$

The transfer matrix will also need to be modified accordingly

$$\bar{\mathbf{H}} = \mathbf{H}\mathbf{L}^{-1} \quad (29)$$

resulting in

$$\min_{\bar{\mathbf{u}}} \|\tilde{\mathbf{y}} - \mathbf{y}\|_2^2$$
$$\text{s.t.} \quad \mathbf{y} = \bar{\mathbf{H}}\bar{\mathbf{u}} \quad (30)$$

The regularization methods are applied to the least squares formulation given in Eq. 30 which replaces Eq. 25. The transcription factor concentrations are calculated from $\bar{\mathbf{u}}$ as described below.

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

$$\bar{\mathbf{u}}_w = \sum_{i=1}^{k} \frac{\bar{\mathbf{u}}_i^T \tilde{\mathbf{y}}}{\bar{w}_i} \bar{\mathbf{v}}_i \quad (31)$$

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

$$\bar{\mathbf{H}} = \bar{\mathbf{U}}\,\bar{\mathbf{W}}\,\bar{\mathbf{V}}^T$$
$$\bar{\mathbf{W}} = \text{diag}\,(\bar{w}_1, \cdots, \bar{w}_{\bar{r}}) \quad (32)$$

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

$$\mathbf{u}_w = \mathbf{L}^{-1}\bar{\mathbf{u}}_w$$
$$C_{TFw}(t_j) = c \frac{u_w(t_j)}{1 - u_w(t_j)} \qquad \forall \; t_0 \leq t_j \leq t_{n-1} \quad (33)$$

Tikhonov Regularization. The Tikhonov regularized solution for $\overline{\mathbf{u}}$ is calculated by solving

$$\min_{\overline{\mathbf{u}}} \|\tilde{\mathbf{y}} - \overline{\mathbf{H}}\,\overline{\mathbf{u}}\|_2^2 + \lambda\|\mathbf{I}_n\overline{\mathbf{u}}\|_2^2 \tag{34}$$

which results in

$$\overline{\mathbf{u}}_\lambda = (\overline{\mathbf{H}}^{\mathrm{T}}\overline{\mathbf{H}} + \lambda\,\mathbf{I}_n)^{-1}(\overline{\mathbf{H}}^{\mathrm{T}}\tilde{\mathbf{y}}) \tag{35}$$

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

$$\begin{aligned} \mathbf{u}_\lambda &= \mathbf{L}^{-1}\overline{\mathbf{u}}_\lambda \\ C_{\mathrm{TF}\lambda}(t_j) &= c\frac{u_\lambda(t_j)}{1-u_\lambda(t_j)} \qquad \forall\ t_0\le t_j\le t_{n-1} \end{aligned} \tag{36}$$

Estimation Error. It is not known a priori which of the two investigated regularization methods is more suitable for the solution of the inverse problem posed in this article. Thus, both regularization methods have been applied and a comparison of the results has been done. The comparison is performed based upon the Relative error (RE)

$$\mathrm{RE}:\ \frac{\|\Omega_{\mathrm{estimated}} - \Omega_{\mathrm{actual}}\|}{\|\Omega_{\mathrm{actual}}\|}\times 100 \tag{37}$$

where Ω is the quantity to be estimated. To ensure a meaningful comparison, the optimal values of the regularization parameters have to be determined from the Picard plot or L-curve for each case.

Regularization with non-negativity constraints

In this subsection, the solution of the inverse problem is evaluated by imposing additional constraints in the optimization formulation (30). For the estimated transcription factor profiles to be physically feasible, the concentrations at each time instant should be non-negative.

$$C_{\mathrm{TF}}(t_j)\ge 0 \qquad \forall\ t_0\le t_j\le t_{n-1} \tag{38}$$

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

$$\begin{aligned} u(t_j)&\ge 0 \quad \forall\ t_0\le t_j\le t_{n-1} \\ u(t_j)&\le 1 \quad \forall\ t_0\le t_j\le t_{n-1} \end{aligned} \tag{39}$$

In formulation (30), $\overline{u}_j$ has been estimated. Given that $\overline{u}_j = u_j - u_{j-1}$ the above constraints can be written as

$$\begin{aligned} \sum_{k=1}^{j}\overline{u}_k &\ge 0\ \ \forall\ \ j = \{0,1,\ldots,n-1\} \\ \sum_{k=1}^{j}\overline{u}_k &\le 1\ \ \forall\ \ j = \{0,1,\ldots,n-1\} \end{aligned} \tag{40}$$

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

Truncated SVD with Additional Constraints.

$$\begin{aligned} \min_{\overline{\mathbf{u}},\mathbf{y}}\ & \|\tilde{\mathbf{y}} - \mathbf{y}\|_2^2 \\ \text{s.t.}\ & \overline{\mathbf{u}} = \overline{\mathbf{K}}\mathbf{y} \\ & \mathbf{R}\,\overline{\mathbf{u}} \ge \mathbf{0} \\ & \mathbf{R}\,\overline{\mathbf{u}} \le \mathbf{1} \end{aligned} \tag{41}$$

where $\mathbf{R}$ is a lower triangular matrix such that the inequality constraint incorporates the constraints in Eq. 40

$$\mathbf{R} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{n\times n} \tag{42}$$

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

$$\overline{\mathbf{K}} = \sum_{i=1}^{k}\frac{\overline{\mathbf{v}}_i\overline{\mathbf{u}}_i^T}{\overline{w}_i} \tag{43}$$

The truncation parameter k is chosen from the Picard condition as described before. The formulation given in (41) is quadratic programming problem. It has been solved by using the solver "quadprog" in MATLAB. Then the transcription factor concentrations are calculated as described in Eq. 33.

Tikhonov regularization with Additional Constraints.

$$\begin{aligned} \min_{\overline{\mathbf{u}}}\ & \|\tilde{\mathbf{y}} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{I}_n\overline{\mathbf{u}}\|_2^2 \\ \text{s.t.}\ & \mathbf{y} = \overline{\mathbf{H}}\overline{\mathbf{u}} \\ & \mathbf{R}\,\overline{\mathbf{u}} \ge \mathbf{0} \\ & \mathbf{R}\,\overline{\mathbf{u}} \le \mathbf{1} \end{aligned} \tag{44}$$

This formulation for Tikhonov Regularization was also solved by using the solver quadprog in MATLAB. Then the transcription factor concentrations are calculated from Eqs. 36. The regularization parameter λ is chosen by examination of the L-curve plotted by using the solution of the above formulation.

Case Studies and Results

In this section, the presented procedure for computing the transcription factor profiles from fluorescence intensity data have been applied to simulated and experimental data. The effects that different regularization techniques and the choice of regularization parameters have on the estimated input profiles are discussed in detail.

Case study 1: simulated data containing Gaussian noise

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

The simulated data were created by assuming a certain profile for the transcription factor concentration and then computing the fluorescence intensity profile resulting from this transcription factor profile by solving Eqs. 1 and 2. Gaussian noise was added to the fluorescent intensity profile to create a more realistic data set. These data are used to solve the inverse problem using both regularization methods. The estimated profiles are compared with the original

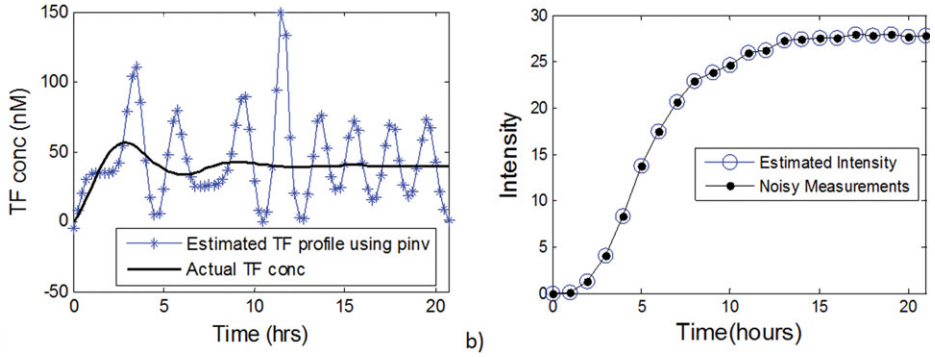

Figure 3. Solution of inverse problem using pinv in MATLAB for measurements containing noise −N(0,0.2) (a) estimated TF profile (b) estimated intensity profile.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

assumed profiles of TF to validate the accuracy of results using the RE. As it is not possible to make comparisons for every potential input profile, the simulated data were computed for a transcription factor profile of the following form

$$C_{\mathrm{TF}}(t) = A(1 - e^{-\alpha t}\cos(t)) \tag{45}$$

where A and α are parameters and $t_0 \leq t \leq t_n$. This profile represents decaying oscillations which is one of the common dynamics exhibited by TF. The value of A and α were assumed to be 40 and 0.3, respectively. Using this form of the transcription factor dynamics, the ODE model was simulated from $t_0 = 0$ to $t_n = 21$ hours. The sampling time for the inputs was chosen to be $\Delta t = 0.25$ hours but the fluorescence intensities data were sampled hourly. This results in a transfer matrix $\mathbf{H} \in R^{21 \times 84}$. Also, note that the parametric form of the transcription factor profile shown in Eq. 45 was assumed solely for data creation purposes and the procedure for solving the inverse problem has no knowledge about this profile.

For illustration purposes, the measurements were initially simulated by adding only a small amount of random Gaussian noise −N(0,0.2). When no regularization has been applied for the solution of this inverse problem, the curve shown in Figure 3a is obtained. This solution was obtained by finding the pseudo-inverse of $\overline{\mathbf{H}}$ using "pinv" in MATLAB. The estimated transcription factor profile is highly erroneous with a RE of 71.5% and the estimated intensity profile, shown in Figure 3b, fits the noisy data well with a RE of 1.04e-13%. Thus, to prevent this over-fitting and filter off the noisy components in the estimated input profile,

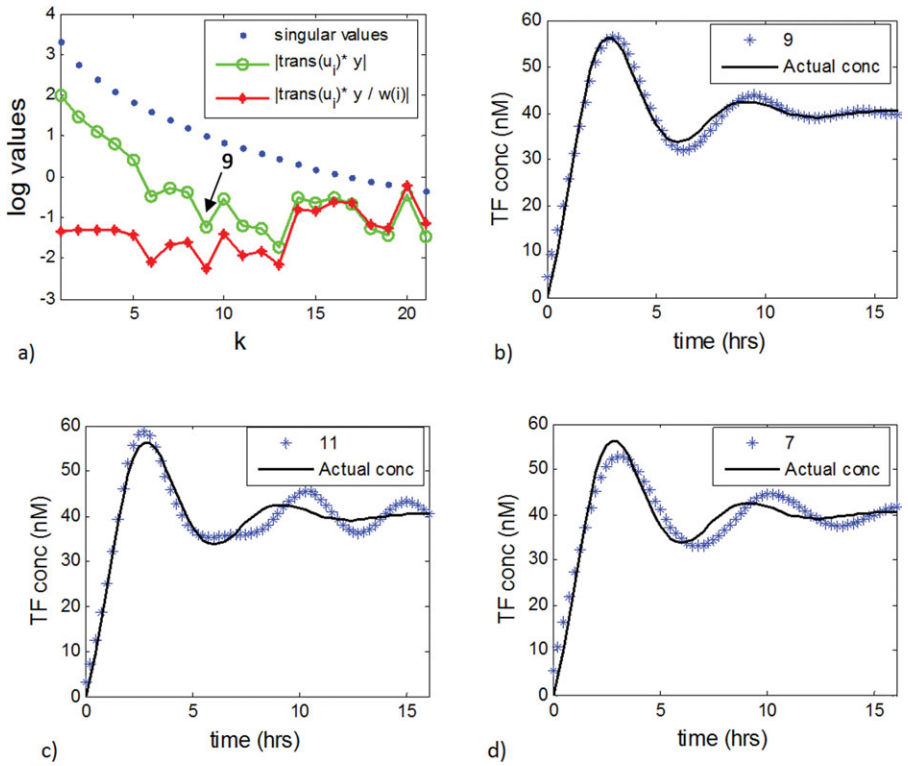

Figure 4. Solution of inverse problem by TSVD for measurements containing noise −N(0,0.2) (a) Picard plot with the optimal truncation parameter of 9 (b) estimated transcription factor profile for $k = 9$ (c) estimated profile with a truncation parameter $k = 11$ (d) estimated profile with a truncation parameter $k = 7$.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

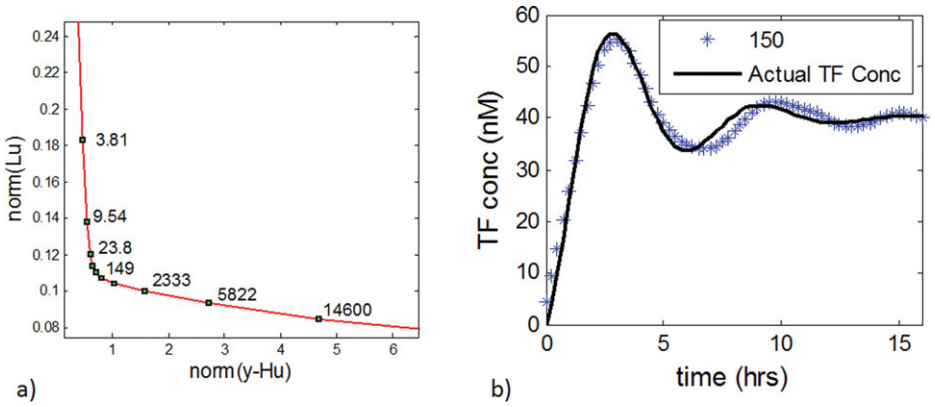

Figure 5. Solution of inverse problem by Tikhonov regularization for measurements containing noise −N(0,0.2) (a) L-curve (b) estimated TF profile using regularization parameter of 150.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

there is a need for regularizing the solution of this inverse problem.

To address over fitting, TSVD was applied to solve the inverse problem. The truncation parameter was found using the Picard condition by plotting the singular values, $\overline{\mathbf{u}}_i^{\mathrm{T}}\tilde{\mathbf{y}}$ and $|\overline{\mathbf{u}}_i^{\mathrm{T}}\tilde{\mathbf{y}}/\overline{w}_i|$ in a semilog plot as shown in Figure 4a. The truncation parameter was chosen to be 9 from the Picard plot as the value of the numerator was found to decay at a rate higher than the singular values until approximately the parameter 9 after which it appears to level off. The estimated profile for the optimal parameter on the basis of the Picard condition is shown in Figure 4b and it results in a RE of 3.68%. For comparison purposes, the estimated profiles using truncation parameters of 11 and 7, instead of the optimal 9, are shown in Figures 4c, d, respectively. It can be seen that the optimal truncation parameter chosen using the Picard condition results in an estimated profile which is a better representation of the actual profile than if other values of the truncation parameter are used. If the truncation parameter is chosen larger than the optimal value, then the computed response is more oscillatory than the real response with a RE of 6.20%. Similarly, if the truncation parameter is chosen to be smaller than desired, then some of the dynamics of the input cannot be correctly reconstructed resulting in a RE of 6.87%. Also, the non-negativity constraints are redundant for this transcription factor profile as the estimated concentrations are safely above zero. So, the problem formulations without the non-negativity constraints were implemented for estimating the transcription factor profiles for the simulated data.

Tikhonov regularization is applied to the same problem for comparison purposes. The regularization parameter, i.e., the value at the corner of the L-curve was found to be ~150 (Figure 5a). Using this regularization parameter, the estimated profile (Figure 5b) gives a low RE of 4.22%. Thus, the results are comparable to what was found using TSVD for regularization for the case where only a small amount of noise was present in the data.

Solution of this inverse problem with a larger noise level has also been performed. One such simulated measurement data set containing random Gaussian noise −N(0,1) is shown in Figure 6. This noise level is more realistic for biological measurements.

When applying truncated SVD to this data set, the optimal truncation parameter was observed to be 6 from the Picard plot (Supporting Information, Figure 1a). The estimated profile shown in Figure 7a has a RE of 11.1%. In the estimated profile, the first peak of the profile is underestimated and the positions of the subsequent peaks are misplaced. Thus, truncated SVD does not perform as well for this example where a larger amount of noise is present in the measurements. Similar observations have been made for even larger noise levels, even though, these results are not shown here.

When Tikhonov regularization was used, the results are also not as good as for the low noise level case, however, they results are better than what was achieved by TSVD for this case. Using a regularization parameter of ~1500, obtained from the L-curve (Supporting Information, Figure 1b), the computed transcription factor profile resulted in the curve shown in Figure 7b which has a RE of 8.33%.

Comparison between truncated SVD and Tikhonov regularization using Monte Carlo simulations

Although the above shown comparisons include a significant amount of detail to explain the methods, they do not allow to draw broad conclusions as only a few specific cases were investigated. This section presents Monte Carlo simulations to provide a detailed comparison between truncated

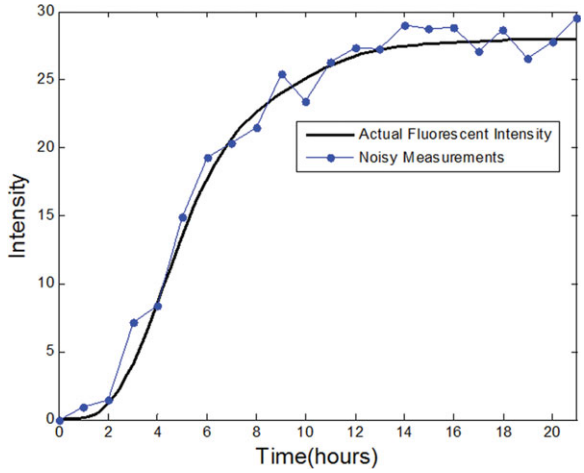

Figure 6. Simulated data containing random Gaussian noise −N(0,1).

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

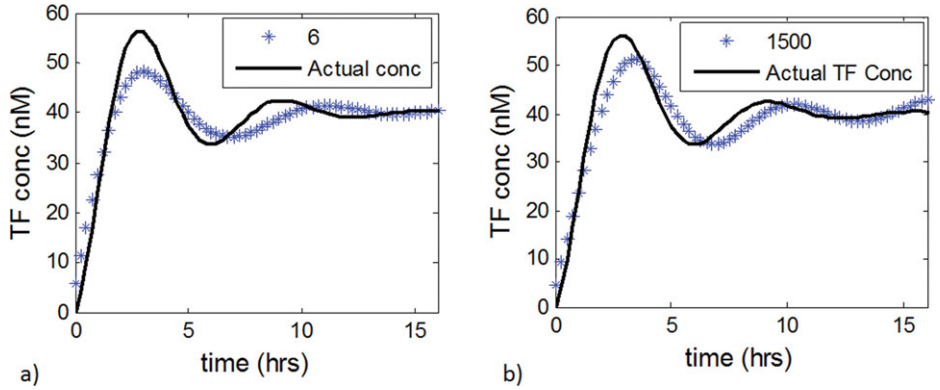

Figure 7. Estimated transcription factor profiles for simulated data containing −N(0,1) a) Using TSVD b) Using Tikhonov regularization.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

SVD and Tikhonov regularization for this inverse problem. Ten thousand data sets were simulated for noise levels of N(0,0.2) and N(0,1) and the inverse problem was solved for each case. The transcription factor profile was assumed to be same as given by Eq. 45. The estimated profiles do not violate the non-negativity constraints and thus these constraints were not used for solving the inverse problem for Monte Carlo simulations. The mean squared error, fitting error, squared bias and variance have been calculated from the estimated input profiles according to the following equations

$$\begin{aligned}
\text{Fitting error} &= tr\{E[(\mathbf{y}-\tilde{\mathbf{y}})(\mathbf{y}-\tilde{\mathbf{y}})^{\mathrm{T}}]\}\\
\text{Mean square error (MSE)} &= tr\{E[(\hat{\mathbf{C}}_{\mathbf{TF}}-\mathbf{C}_{\mathbf{TF}})(\hat{\mathbf{C}}_{\mathbf{TF}}-\mathbf{C}_{\mathbf{TF}})^{\mathrm{T}}]\}\\
\text{Bias}^2 &= tr\{(E[\hat{\mathbf{C}}_{\mathbf{TF}}]-\mathbf{C}_{\mathbf{TF}})(E[\hat{\mathbf{C}}_{\mathbf{TF}}]-\mathbf{C}_{\mathbf{TF}})^{\mathrm{T}}\}\\
\text{Variance} &= tr\{E[(\hat{\mathbf{C}}_{\mathrm{TF}}-E[\hat{\mathbf{C}}_{\mathbf{TF}}])(\hat{\mathbf{C}}_{\mathbf{TF}}-E[\hat{\mathbf{C}}_{\mathbf{TF}}])^{\mathrm{T}}]\} \quad (46)
\end{aligned}$$

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

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

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

Furthermore, the error bounds of the estimated transcription factor concentrations have been calculated to illustrate the variability in the estimated profiles. The following

Table 1. Results from Monte Carlo Simulations of 10,000 Simulated Data Sets Containing Noise −N(0,0.2)

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

Results from Monte Carlo Simulations of 10,000 Simulated Data Sets Containing Noise −N(0,0.2)

Table 2. Results from Monte Carlo Simulations of 10,000 Simulated Data Sets Containing Noise −N(0,1)

Parameter Used	MSE	Fitting Error	Bias^2	Variance
TSVD				
3	8028.72	63.79	7845.07	190.78
4	3313.80	24.60	2980.80	336.96
5	**1754.91**	**16.73**	**965.97**	**787.18**
6	2155.98	15.39	643.19	1515.10
7	3431.29	14.23	524.23	2928.95
Tikhonov				
375	2919.28	12.78	283.89	2641.68
938	1839.01	14.32	504.73	1332.10
2344	**1578.65**	**16.81**	**879.46**	**697.19**
5859	1954.84	22.63	1576.03	374.96
14,648	3004.21	37.75	2793.19	211.24
'pinv' solution	2.81E + 13	2.40E − 26	4.01E + 11	2.77E + 13

Results from Monte Carlo Simulations of 10,000 Simulated Data Sets Containing Noise −N(0,1)

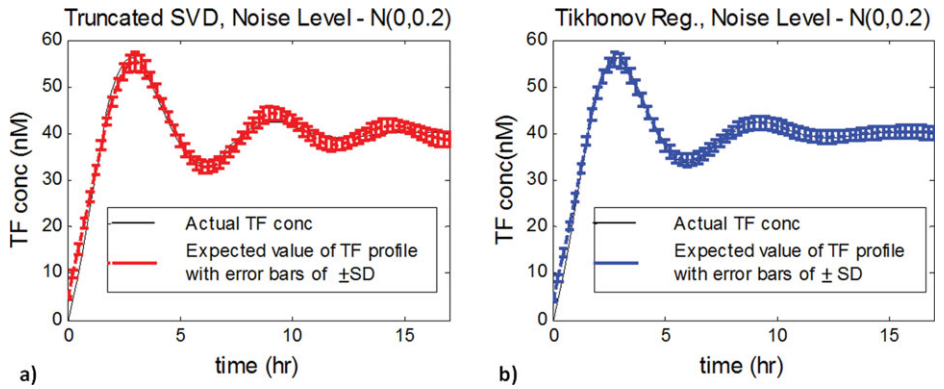

Figure 8. Expected value of TF profile for noise level of N(0,0.2) with error bars representing ± SD using (a) truncated SVD (b) Tikhonov regularization.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

equation is used for calculating the standard deviation (SD) of the transcription factor concentrations at each time point

$$e(t_i) = (\text{diag}\{E[(\hat{\mathbf{C}}_{\mathbf{TF}} - E[\hat{\mathbf{C}}_{\mathbf{TF}}])(\hat{\mathbf{C}}_{\mathbf{TF}} - E[\hat{\mathbf{C}}_{\mathbf{TF}}])^{\mathrm{T}}]\}_i)^{\frac{1}{2}} \quad \forall\, t_i \in [t_o, t_n] \tag{47}$$

where $e(t_i)$ refers to the standard deviation for the transcription factor concentration estimated at time t_i and $\text{diag}\{\cdot\}_i$ refers to the i^{th} element of the diagonal of the matrix. The estimated profiles for the noise level −N(0,0.2) and the error bars are shown in Figure 8. The length of the error bars is 2 $e(t_i)$ and the expected profile and the error bars are corresponding to the regularization parameter that resulted in the lowest MSE for each regularization method. For the noise level −N(0,0.2), the standard deviations were within 4.5% and 3.7% of the expected value of the transcription factor profiles calculated using TSVD and Tikhonov regularization, respectively. This calculation excludes the errors obtained for the last few hours of the data which are significantly larger than for the rest of the profile. The reason for this is that transcription, translation, and post-translational modification requires a certain amount of time to take place and transcription factor concentrations that are computed towards the end of an experiment would largely be affected by fluorescence intensities which are occurring after an experiment has concluded. Therefore, the TF concentrations for the last few hours cannot be accurately estimated from the intensity data available for the same time range.

Case study 2: application to experimental data using non-negativity constraints

The previous case study illustrated in detail the effects that different choices of regularization parameters have on the solution of the inverse problem using simulated data. However, the most important test is to apply the procedure to experimental data to determine if the procedure will return satisfactory results. The experimental data are available in the form of a series of fluorescent images of a GFP reporter system (Figure 9a) taken during the course of the experiment. The images have been analyzed to remove noisy pixels and obtain a time-dependent mean fluorescence intensity profile.[11] These data are used to estimate the dynamic profiles of transcription factor by solving the inverse problem.

For this case study, experimental data were obtained for the transcription factor STAT3 by continuously stimulating liver cells with 100 ng/ml of IL-6 using a previously developed procedure.[24] The fluorescent microscopy images were taken every 45 min for a period of 22 hours at multiple positions in the well. The mean fluorescence intensity of the images at each time instant were calculated[11] and are shown in Figure 9b. The shown profile was used to solve the inverse problem to obtain profiles for STAT3. It can be seen

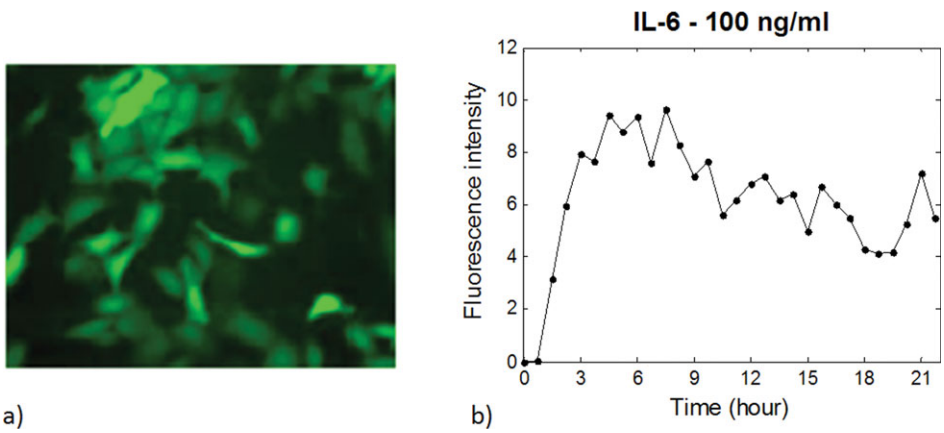

Figure 9. (a) Sample image obtained from fluorescence microscopy of a GFP reporter system (b) fluorescence intensity profile obtained for IL-6-STAT3 system.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

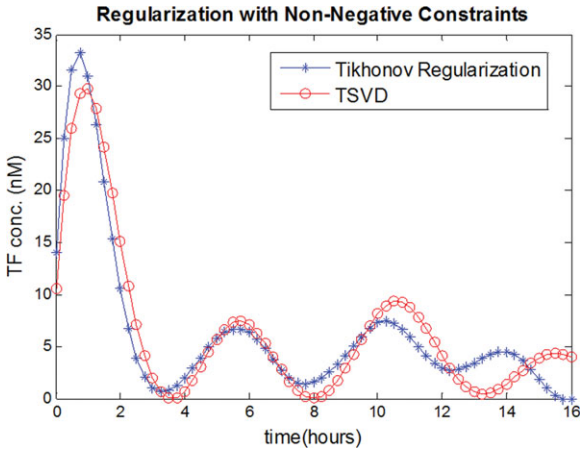

Figure 10. Estimated STAT3 profiles from Tikhonov regularization and truncated SVD.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

from these data that the experimental measurements contain a significant amount of noise and regularization should be applied to prevent over-fitting.

It is known that the initial dynamics of the transcription factor STAT3 shows a rapid increase followed by a steep decrease. The reason for this is that cytoplasmic STAT3 is activated and translocates to the nucleus after few minutes of stimulation with IL-6.[32–34] Thus, the input was discretized for a sampling time of 15 min to be able to infer the initial dynamics, but the sampling time was also not chosen to be too small to ensure that the inverse problem does not grow very large in size as output data were available for a time period of 22 hours. In addition, the STAT3 concentrations cannot be negative, the two regularization techniques were applied along with non-negativity constraints. If non-negativity constraints are not used, the estimated STAT3 profiles using both TSVD and Tikhonov regularization attain negative values at a number of time points (Supporting Information, Figure 2).

The optimal truncation parameters for this data were found to be 9 for TSVD and 350 for Tikhonov regularization from the Picard plot and L-curve, respectively (Supporting Information, Figure 3). It can be seen from the estimated results shown in Figure 10 that the STAT3 profile is oscillatory in nature with a large initial peak followed by a smaller peak at around 5–6 hours and potentially another peak at around 10–11 hours. These results are consistent with Western blot data as well as simulation results of the IL-6 signal transduction pathway given in the literature.[34,35] Although the solution of the transcription factor profile involving TSVD suggests similar locations of the peaks as the solution involving Tikhonov regularization, the TSVD solution appears to be more oscillatory.

Furthermore, the results returned by TSVD suggest that the ratio of the peak concentration of STAT3 between the first and the second peak is $\sim$4 whereas the ratio between the concentrations of the first and the second peak is $\sim$5 when Tikhonov regularization is used. Simulation studies of the JAK-STAT pathway[36] suggest that the ratio of the first and the second peak is $\sim$5 which is more consistent with results returned by Tikhonov regularization. Moreover, the concentration profiles of the TF estimated using TSVD are almost zero at a certain time points which goes against what one would expect for this system.

Conclusions

This article presented a general method for extracting transcription factor profiles from fluorescence intensity profiles. This technique involves formulating and solving an inverse problem which directly relates the output of a GFP reporter system, i.e., the fluorescent intensity, to the input which is a function of the transcription factor concentration. The procedure used in this work places no restrictions on the shape of the input, unlike previous studies, where the transcription rates or transcription factor profiles had to be of a certain nature.[5,11] This was achieved by discretizing the input at certain points in time and then solving an inverse problem which computes the transcription factor concentration at each discrete time point.

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

Acknowledgments

The authors acknowledge the fluorescence microscopy images provided by Dr. Arul Jayaraman and Mr. Colby Moya. The authors gratefully acknowledge the partial financial support from the National Science Foundation (Grant CBET#0941313).

Notation

GFP = green fluorescent protein
GLS = generalized least squares
IL-6 = interleukin-6
MSE = mean squared error
nM = nano molar
ODE = ordinary differential equation
RE = relative error
SD = standard deviation
STAT3 = signal transducer and activator of transcription 3
SVD = singular value decomposition
TF = transcription factor
TSVD = truncated singular value decomposition

Literature Cited

1. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*. 2004;431:308–312.
2. Jothi R, Balaji S, Wuster A, et al. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Mol Syst Biol*. 2009;5:294.
3. Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC. Green fluorescent protein as a marker for gene expression. *Science*. 1994;263:802–805.
4. Roessel Pv, Brand AH. Imaging into the future: visualizing gene expression and protein interactions with fluorescent proteins. *Nat Cell Biol*. 2002;4:15–20.

5. Finkenstädt B, Heron EA, Komorowski M, et al. Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics*. 2008;24:2901–2907.
6. Wang X, Errede B, Elston TC. Mathematical analysis and quantification of fluorescent proteins as transcriptional reporters. *Biophys J*. 2008;94:2017–2026.
7. Leveau JHJ, Lindow SE. Predictive and interpretive simulation of green fluorescent protein expression in reporter bacteria. *J Bacteriol*. 2001;183:6752–6762.
8. De Jong H, Ranquet C, Ropers D, Pinel C, Geiselmann J. Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. *BMC Syst Biol*. 2010;4:55.
9. Subramanian S, Srienc F. Quantitative analysis of transient gene expression in mammalian cells using the green fluorescent protein. *J. Biotechnol*. 1996;49:137–151.
10. Ronen M, Rosenberg R, Shraiman B, Alon U. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci USA*. 2002;99:10555–10560.
11. Huang Z, Senocak F, Jayaraman A, Hahn J. Integrated modeling and experimental approach for determining transcription factor profiles from fluorescent reporter data. *BMC Syst. Biol*. 2008;2:64.
12. Huang Z, Chu Y, Cunha B, Hahn J. Generalisation of a procedure for computing transcription factor profiles. *IET Syst Biol*. 2010;4:108–118.
13. Dössel O. Inverse problem of electro- and magnetocardiography: review and recent progress. *Int J Bioelectromagn*. 2000;2.
14. Zhdanov MS. *Geophysical Inverse Theory and Regularization Problems*, Vol. 36. Elsevier Science B.V, Amsterdam, The Netherlands, 1st ed., 2002.
15. Borcea L. Electrical impedance tomography. *Inverse Prob*. 2002;18:99–136.
16. Hansen PC. Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank. *SIAM J. Sci. Comput*. 1990;11:503–518.
17. Hansen PC. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev*. 1992;34:561–580.
18. Aster RC, Borchers B, Thurber C. *Parameter Estimation and Inverse Problems*. Academic Press, Oxford, UK, 2005.
19. Fierro RD, Golub GH, Hansen PC, O'Leary DP. Regularization by truncated total least squares. *SIAM J Sci Comput*. 1997;18:1223.
20. Vogel CR. *Computational Methods for Inverse Problems*. SIAM, Philadelphia, PA, 2002.
21. Hansen PC. *Discrete Inverse Problems: Insight and Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2010.
22. Golub GH, Hansen PC, O'Leary DP. Tikhonov regularization and total least squares. *SIAM J Matrix Anal Appl*. 2000;21:185–194.
23. Shou G, Xia L, Jiang M, Wei Q, Liu F, Crozier S. Truncated total least squares: a new regularization method for the solution of ECG inverse problems. *IEEE Trans Biomed Eng*. 2008;55:1327–1335.
24. Moya C, Huang Z, Cheng P, Jayaraman A, Hahn J. Investigation of IL-6 and IL-10 signalling via mathematical modelling. *IET Syst Biol*. 2009;5:15–26.
25. Tarantola A. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2005.
26. Verkruysse W, Majaron B, Choi B, Nelson JS. Combining singular value decomposition and a non-negative constraint in a hybrid method for photothermal depth profiling. *Rev Sci Instrum*. 2005;76:024301.
27. Zhu X, Shen J, Liu W, Sun X, Wang Y. Nonnegative least-squares truncated singular value decomposition to particle size distribution inversion from dynamic light scattering data. *Appl Opt*. 2010;49:6591–6596.
28. Villiers GD, McNally B, Pike E. Positive solutions to linear inverse problems. *Inverse Probl*. 1999;15:615.
29. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed. Springer, New York, 2009.
30. Rojas M, Steihaug T. An interior-point trust-region-based method for large-scale non-negative regularization. *Inverse Probl*. 2002;18:1291.
31. Chiang Y-W, Borbat PP, Freed JH. Maximum entropy: a complement to Tikhonov regularization for determination of pair distance distributions by pulsed ESR. *J Magn Reson*. 2005;177:184–196.
32. Watanabe K, Saito K, Kinjo M, et al. Molecular dynamics of STAT3 on IL-6 signaling pathway in living cells. *Biochem Biophys Res Commun*. 2004;324:1264–1273.
33. Kretzschmar AK, Dinger MC, Henze C, Brocke-Heidrich K, Horn F. Analysis of Stat3 (signal transducer and activator of transcription 3) dimerization by fluorescence resonance energy transfer in living cells. *Biochem J*. 2004;377:289.
34. Singh A, Jayaraman A, Hahn J. Modeling regulatory mechanisms in IL-6 signal transduction in hepatocytes. *Biotechnol Bioeng*. 2006;95:850–862.
35. Fischer P, Lehmann U, Sobota RM, et al. The role of the inhibitors of interleukin-6 signal transduction SHP2 and SOCS3 for desensitization of interleukin-6 signalling. *Biochem J*. 2004;378:449–460.
36. Yamada S, Shiono S, Joo A, Yoshimura A. Control mechanism of JAK/STAT signal transduction pathway. *FEBS Lett*. 2003;534:190–196.

Appendix: Evaluation of the Transfer Matrix H

The integrals in Eq. 22 are evaluated using eigenvalue decomposition of $\mathbf{A}$

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

where

$\Lambda = \begin{bmatrix} \Lambda_1 & & 0 \\ & \ddots & \\ 0 & & \Lambda_q \end{bmatrix}$; Λ_i's are the eigenvalues of the $\mathbf{A}$ matrix and the columns of the Σ matrix are the eigenvectors.

The integrals can be evaluated as

$$\int_a^b e^{\mathbf{A}(T-\tau)}d\tau = \int_b^a e^{\mathbf{A}(T-\tau)}d(T-\tau)$$
$$= \int_{T-b}^{T-a} e^{\mathbf{A}\tau}d\tau = \Sigma\left(\int_{T-b}^{T-a} e^{\Lambda\tau}d\tau\right)\Sigma^{-1}$$
$$= \Sigma\begin{bmatrix} \int_{T-b}^{T-a} e^{\Lambda_1\tau}d\tau & & 0 \\ & \ddots & \\ 0 & & \int_{T-b}^{T-a} e^{\Lambda_q\tau}d\tau \end{bmatrix}\Sigma^{-1}$$
$$= \Sigma\begin{bmatrix} \frac{1}{\Lambda_1}\left(e^{\Lambda_1(T-a)} - e^{\Lambda_1(T-b)}\right) & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\Lambda_q}\left(e^{\Lambda_q(T-a)} - e^{\Lambda_q(T-b)}\right) \end{bmatrix}\Sigma^{-1}$$

Manuscript received Sept. 15, 2011, and revision received Feb. 10, 2012.