Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Optimal design of non-equilibrium experiments for genetic network interrogation

Kaska Adoteye, H.T. Banks *, Kevin B. Flores
Department of Mathematics, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, United States

A R T I C L E   I N F O

Article history:
Received 21 September 2014
Accepted 21 September 2014
Available online 28 September 2014

Keywords:
Optimal experimental design
Inverse problem
Uncertainty analysis
Brome mosaic virus
Synthetic biology
Genetic networks

A B S T R A C T

Many experimental systems in biology, especially synthetic gene networks, are amenable to perturbations that are controlled by the experimenter. We developed an optimal design algorithm that calculates optimal observation times in conjunction with optimal experimental perturbations in order to maximize the amount of information gained from longitudinal data derived from such experiments. We applied the algorithm to a validated model of a synthetic Brome Mosaic Virus (BMV) gene network and found that optimizing experimental perturbations may substantially decrease uncertainty in estimating BMV model parameters.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Recent efforts in modeling the host immune response to HIV infection have illuminated the relationship between perturbations that drive biological systems away from equilibrium and information content in data measured from such systems [1,2]. For example, the HIV model developed by Banks, et al. [3,4] describes how anti-retroviral therapy (ART) drives viral load in patients toward an equilibrium level that is undetectable, even by ultra-sensitive assays. When ART is interrupted, e.g., due to patient non-adherence, the HIV model converges toward an equilibrium with high viral load. Indeed, these are the dynamics observed in clinical patient data [4]. Banks, et al. fit their HIV model to clinical patient data and exhibited that the number of HIV model parameters that could be estimated with high statistical confidence increased with the number of treatment interruptions [2]. Thus, non-equilibrium dynamics, induced by ART perturbations, increased the data information content as calculated through asymptotic standard errors for estimated model parameters.

We hypothesized that this positive relationship between information content and system perturbations may exist for more general mathematical models and in particular for models describing biological networks. To investigate this relationship, we employed an optimal experimental design theory framework [5–7] to develop an algorithm that minimizes parameter standard errors by choosing optimal perturbations to experimental inputs. Specifically, we describe how the algorithm for optimizing selection of observation times can be extended to include optimization of experimentally controlled perturbations in order to produce data sets with maximal information content. Although we do not propose intentional perturbations in a clinical setting with patients, such a framework could be useful for gaining information from in vitro experiments where there may exist limitations on the number of observable states and observation times.

A particularly useful application of our algorithm involves interrogation of genetic networks. Data from genetic networks can be collected by measuring longitudinal gene expression, either pre- or post-translational, from in vitro cell lines.

* Corresponding author.
E-mail address: htbanks@eos.ncsu.edu (H.T. Banks).

http://dx.doi.org/10.1016/j.aml.2014.09.013
0893-9659/© 2014 Elsevier Ltd. All rights reserved.
2.1. Mathematical models, statistical models, and parameter uncertainty quantification

Estimating model parameters thereby dramatically decreases standard errors for estimated model parameters, i.e., reduces dramatically the uncertainty in the experimentally controlled inputs (chemicals) for the BMV system can lead to more informative experiments, and thereby reduces uncertainty in estimated model parameters.

2.2. Optimal design measures

Importantly, there are also several methods for experimentally perturbing in vitro gene expression at the pre- and post-transcriptional levels [8,9]. We recently estimated kinetic parameters for a model of a synthetically constructed gene network for the recruitment module of the Brome Mosaic Virus replication cycle [10,11]. In the BMV synthetic system, gene expression is tuned by the concentration of experimentally controlled chemicals. Here, we report how optimization of the experimentally controlled inputs (chemicals) for the BMV system can lead to more informative experiments, and thereby dramatically decrease standard errors for estimated model parameters, i.e., reduce dramatically the uncertainty in estimating model parameters.

2. Data and methods

2.1. Mathematical models, statistical models, and parameter uncertainty quantification

In this note, we formulate an optimal design framework for experimental systems with a scalar time-dependent input \( b(t) \). In practice, \( b(t) \) is assumed to be known since it is controlled by the experimenter.

The mathematical model we consider is

\[
\begin{align*}
\frac{d\bar{x}}{dt} &= \tilde{g}(t, \bar{x}(t; \tilde{\theta}), \tilde{q}, b(t)), \quad t \in [t_0, t_f], \\
\bar{x}(t_0, \tilde{\theta}) &= \bar{x}_0, \\
\tilde{f}(t, \tilde{\theta}, b(t)) &= C_{\theta}(t, \tilde{\theta}, b(t))
\end{align*}
\]

where \( \bar{x}(t, \tilde{\theta}) \) is the vector of state variables of the system generated using a parameter vector \( \tilde{\theta} = (\bar{x}_0, \tilde{q}) \in \mathbb{R}^p, p = N + r \), that contains \( N \) initial values and \( r \) system parameters listed in \( \tilde{q} \). The map \( \tilde{g} : \mathbb{R}^{1+N+r} \rightarrow \mathbb{R}^N \) has the corresponding observation process \( \tilde{f}(t, \tilde{\theta}, b(t)) = C_{\theta}(t, \tilde{\theta}, b(t)) \) with observation operator \( C \) that connects the model solution to observed data. Here, \( C \) is a \( K \times N \) matrix, where \( K \leq N \) is allowed. The times \( t_0 \) and \( t_f \) are initial and final experiment times, respectively. To illustrate the inverse problem methodology, we use a constant i.i.d statistical error model, although more general error formulations can be readily derived and treated. Further statistical details, including a description of the associated \( K \times K \) covariance matrix \( V_0 \), can be found in [12]. In this work we consider the simple case where \( b(t) \) can be described as a binary vector \( \tilde{b} \) of length \( H \), with values in \{0, 1\} that represent whether the experimental input is on or off in the time intervals \([t_{i-1}, t_i]\), \( i = 1, \ldots, H \).

For a given member \( \theta_k \) of the estimated parameter vector \( \tilde{\theta} \) the standard error (SE) is computed by standard methods from asymptotic theory. For Tables 1 and 2, the normalized standard error (NSE) is defined as \( \frac{\theta_k - 1.96SE_k}{SE_k} \); the 95% confidence interval (CI) is given by [\( \theta_k - 1.96SE_k \), \( \theta_k + 1.96SE_k \)] (see [12] for details).

2.2. Optimal design measures

We follow the optimal design formulation using the Generalized Fisher Information Matrix [5–7]. Let \( \mathcal{P}_1(t_0, t_f) \) denote the set of all bounded distributions on the interval \([t_0, t_f]\). Let \( B = \mathbb{R}^H_2 \) be the set of binary vectors \( \tilde{b} \) of length \( H \) that represent the input perturbation \( b(t) \). Let \( \mathcal{P}_2(B) \) represent the set of all bounded distributions \( \mathcal{P}_2(b) \) on \( B \). Then the GFIM may be written as

\[
\mathcal{F}(\mathcal{P}_1, \mathcal{P}_2, \tilde{\theta}_0) = \int_{t_0}^{t_f} \int_{\mathbb{R}^H_2} \nabla^2_{\tilde{\theta}_0} \tilde{f}(t, \tilde{\theta}_0, b(t)) \left( V_0^{-1}(t) \right) \nabla_{\tilde{\theta}_0} \tilde{f}(t, \tilde{\theta}_0, b(t)) \cdot d\mathcal{P}_2(b) \cdot d\mathcal{P}_1(t).
\]
We consider the case of observations collected at discrete times where we choose a set of $n$ time points $\tau = \{t_j\}, j = 1, 2, \ldots, n$, and $t_0 = t_1 < t_2 < \cdots < t_n = t_f$. The corresponding discrete $p \times p$ Fisher information matrix (FIM) for a discrete input $b$ measured at discrete times $\tau$ is

$$F(\tau, b, \theta_0) = \sum_{j=1}^{n} \nabla_{\tilde{\theta}_0} f(t_j, \tilde{\theta}_0, b(t_j)) \left( V_{\tilde{\theta}_0}^{-1}(t_j) \right) \nabla_{\tilde{\theta}_0} f(t_j, \tilde{\theta}_0, b(t_j)).$$

(3)

Methods for calculating the sensitivities $\nabla_{\tilde{\theta}} f(t, \tilde{\theta}, b(t))$ for delay differential equations, such as the model we consider below, are described in [13]. The choice of optimal design criteria is given by the minimization of a functional $\mathcal{J}(\mathcal{F}) : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^+$; a description of SE-, D-, and E-optimal design criteria can be found in [6].

2.3. Non-equilibrium experimental design algorithm

The algorithm is initialized with an initial experimental design consisting of an ordered set of $n$ sampling times, $\tau$, and a vector $b$ of $H$ ones for the experimental input $b(t)$. This initial design represents the unoptimized, or naive, experimental design in which the input is always on. Calculating the optimal $(\tau, b)$-pair requires a computationally demanding nonlinear optimization of $n$ time points and $2^H$ possible input vectors (a total of $n + 2^H$ dimensions). We instead iteratively solve the set of coupled equations

$$\hat{b}^* = \text{argmin}_{\{b|\tau_{b} \in \mathcal{P}_2(b)\}} \mathcal{J}(F(\tau^*, \hat{b}, \tilde{\theta}_0))$$

(4)

$$\tau^* = \text{argmin}_{\{\tau|\tau_{b} \in \mathcal{P}_1(\tau_{b} = t_f)\}} \mathcal{J}(F(\tau, \hat{b}^*, \tilde{\theta}_0))$$

(5)

where $\mathcal{J}$ represents the SE-, D-, or E-optimal design criteria.

3. Results

3.1. A gene network model for RNA3 recruitment in Brome Mosaic Virus

We applied our optimal design framework to the following previously validated model [10] of RNA3 recruitment in the Brome Mosaic Virus (BMV) replication cycle.

$$\frac{dx}{dt} = b(t) \frac{r_x}{1 + Ae^{-x(t)}} - d_x x(t)$$

$$\frac{dy}{dt} = r_y - d_y y(t) - mx(t) - sy(t)$$

$$\frac{dz}{dt} = mx(t) - sy(t) - d_z z(t).$$

(6)

This model was developed to investigate the recruitment process in the replication cycle of BMV, a positive strand RNA virus. This replication cycle is highly conserved across positive strand RNA viruses, such as Severe Acute Respiratory Syndrome (SARS) and Hepatitis C, and the BMV system has been used to gain insights into interactions of the virus with host factors [11,14,15]. Briefly, the mathematical model describes the interaction between Protein 1a ($x(t)$) and RNA3 in the unstabilized ($y(t)$) and stabilized ($z(t)$) forms; for an in-depth description see [10]. The levels of Protein 1a ($x(t)$) and total RNA3 ($y(t) + z(t)$) are estimated prior to estimating parameters for RNA3. Thus, below we treat Protein 1a parameters as known and only estimate the RNA3 parameters: $r_y, d_y, m, d_z$, and the time delay $s$. The values we used for estimated model parameters and the variance of the statistical error model can be found in [10].

| Parameter | $r_x$ | $d_x$ | $d_y$ | $m$ | $s$ |
|-----------|-------|-------|-------|-----|-----|
| Estimate  | 31.641| 0.7562| 0.3139| 0.5557| 1.2374|
| NSE (D)  | 0.0852| 0.1052| 0.1049| 0.3210| 0.0583|
| 95% CI (D)| (26.3558,36.9262) | (0.60023,0.91217) | (0.24835,0.37845) | (0.20603,0.90537) | (1.0598,1.379) |
| NSE (E)  | 0.0541| 1.5602| 0.0901| 1.0197| 0.9503|
| 95% CI (E)| (28.2845,34.9975) | (−1.5563,0.0687) | (0.25845,0.36935) | (−0.55494,1.6663) | (−1.0676,3.5424) |
| NSE (SE) | 0.0599| 0.0840| 0.0701| 0.3173| 0.0665|
| 95% CI (SE)| (27.9255,35.3565) | (0.63163,0.88077) | (0.27075,0.35705) | (0.21005,0.90135) | (1.0761,1.3987) |
The model was developed to describe an experiment performed in synthetic yeast that contained plasmids for RNA3 and protein 1a whose expression is controlled by the concentration of copper and galactose, respectively. Data were collected under equilibrium experimental conditions, i.e., both copper and galactose were given at constant concentrations and the biological system described by Eq. (6) converged toward a constant equilibrium. Importantly, previous data did not support a high confidence in the estimation of several RNA3 parameters [10]. We subsequently hypothesized that creating a non-equilibrium experiment in which the galactose input is allowed to vary on or off, and copper is held constant, would lead to increased statistical confidence in RNA3 parameter estimates. The function $b(t)$ represents the input and, below, we discretize $b(t)$ into an $H = 10$ dimensional binary vector, $\vec{b}$.

3.1.1. Naive experimental design and non-iterative algorithm results

We first compared results from the unchanged naive experimental design (such as used in [10]) to a non-iterative version of the optimal design algorithm described above, i.e., optimizing only the observation times ($\tau$) or the input $\vec{b}$. For each case, we considered a scenario with 27 experimental observation times of total RNA3 over 26 h, where the initial and final times were fixed at $t = 0$ h to $t = 26$ h, respectively (Fig. 1 and Fig. 2). Only results from SE-optimal design criteria were plotted in Fig. 1, since this criterion, unsurprisingly, results in the lowest standard errors for each parameter. We also consider the simple case in which the time intervals over which $\vec{b}$ is discretized, $[t_{bi-1}, t_{bi}]$, $i = 1, \ldots, H$, are of equal length.

We found that optimizing the input $\vec{b}$ with the SE-optimal design criteria resulted in lower normalized standard errors (NSE) for each parameter as compared to optimizing the time points $\tau$ or the naive experimental design (Table 1). Among optimizations of $\vec{b}$, SE-optimal design criteria outperformed the D- and E- criteria when considering the overall sum of the NSEs (see Fig. 3).

3.1.2. Iterative algorithm results

We next compared results from SE-, D-, and E- optimal design criteria when iterating between Eqs. (4) and (5) (see Fig. 2). We found that the effectiveness in using the algorithm allowed the use of less observation time points; hence in the results below we used 14 observation time points instead of 27. Overall, the iterative algorithm outputs an experimental design which may result in significantly lower standard errors for all parameter estimates as compared to the naive experimental design regardless of the optimal design criteria choice (Fig. 3 (left), zero iterations = naive experimental design). Between the optimal design criterion, the SE-optimal design resulted in the lowest sum of NSEs, followed by D-optimal and E-optimal designs, although we note that there was variability in this comparison for each individual parameter (Table 2).

4. Discussion

Overall, our results suggest that experimental input manipulation can produce non-equilibrium system dynamics, leading to a greater information content in collected data. Taking the non-iterative algorithm results together with the iterative
algorithm results, our findings suggest that input manipulation is a more powerful tool for reducing standard errors in parameter estimates than optimizing observation times for the BMV system. For example, optimizing only the observation times still resulted in unreasonably large confidence intervals for the parameter $m$, whereas optimizing only the experimental input resulted in acceptably narrow confidence intervals for $m$, as well as extremely narrow confidence intervals for all other parameters regardless of the choice of optimal design criteria (Table 1).

In future investigations, we will extend the BMV model to consider multiple time-dependent inputs for both Protein 1a and RNA3, since they are controlled separately by the concentration of galactose and copper, respectively. We postulate that, in general, lower standard errors can be achieved when a greater number of system variables are manipulated with experimentally controlled inputs. In addition, we are currently exploring the use of the iterative algorithm (Eqs. (4), (5)) in other genetic network systems that approach a periodic equilibrium to test whether the structure of the $\omega$-limit set affects algorithm convergence.

**Acknowledgments**

This research was supported in part by Grant number NIAID R01AI071915-10 from the National Institute of Allergy and Infectious Diseases, in part by the Air Force Office of Scientific Research under Grant number AFOSR FA9550-12-1-0188, and in part by the National Science Foundation under Research Training Grant (RTG) DMS-1246991 and Grant number DMS-0946431.
References

[1] R. Baraldi, K. Cross, C. McChesney, L. Poag, E. Thorpe, K.B. Flores, H.T. Banks, Uncertainty quantification for a model of HIV-1 patient response to antiretroviral therapy interruptions, in: Proceedings of the American Control Conference, 2014, pp. 2753–2758.
[2] H.T. Banks, R. Baraldi, K. Cross, K.B. Flores, C. McChesney, L. Poag, E. Thorpe, Uncertainty quantification in modeling HIV viral mechanics, CRSC-TR13-16, Center for Research in Scientific Computation, North Carolina State University, 2013, Math. Biosci. Eng. submitted for publication.
[3] B.M. Adams, H.T. Banks, M. Davidian, E.S. Rosenberg, Model fitting and prediction with HIV treatment interruption data, Bull. Math. Biol. 69 (2007) 563–584.
[4] H.T. Banks, M. Davidian, S. Hu, G.M. Kepler, E.S. Rosenberg, Modeling HIV immune response and validation with clinical data, J. Biol. Dyn. 2 (2008) 357–385.
[5] H.T. Banks, S. Dediu, S.L. Ernstberger, F. Kappel, Generalized sensitivities and optimal experimental design, J. Inverse Ill-Posed Probl. 18 (2010) 25–83.
[6] H.T. Banks, K. Holm, F. Kappel, Comparison of optimal design methods in inverse problems, Inverse Problems 27 (2011) 075002.
[7] H.T. Banks, K.L. Rehm, Experimental design for distributed parameter vector systems, Appl. Math. Lett. 26 (1) (2013) 10–14.
[8] K.F. Murphy, R.M. Adams, X. Wang, G. Balazsi, J.J. Collins, Tuning and controlling gene expression noise in synthetic gene networks, Nucleic Acids Res. 38 (8) (2010) 2712–2726.
[9] G. Chalancon, C.N.J. Ravarani, S. Balaji, A. Martinez-Arias, L. Aravind, R. Jothi, M.M. Babu, Interplay between gene expression noise and regulatory network architecture, Trends Genet. 28 (5) (2012) 221–232.
[10] H.T. Banks, K.B. Flores, T. Huffman, K. Link, J. Nardini, L. Poag, B. Blasco, J. Jungfleisch, J. Diez, A mathematical model of RNA3 recruitment in the Brome Mosaic Virus replication cycle, Int. J. Pure Appl. Math. 89 (2) (2013) 251–274.
[11] I. Alves-Rodrigues, R.P. Galao, A. Meyerhans, J. Diez, Saccharomyces cerevisiae: a useful model host to study fundamental biology of viral replication, Virus Res. 120 (2006) 49–56.
[12] H.T. Banks, S. Hu, W.C. Thompson, Modeling and Inverse Problems in the Presence of Uncertainty, CRC Press, Boca Raton, 2014.
[13] H.T. Banks, D. Robbins, K.L. Sutton, Theoretical foundations for traditional and generalized sensitivity functions for delay differential equations, Math. Biosci. Eng. 10 (2013) 1301–1333.
[14] M. Schwartz, J. Chen, M. Janda, M. Sullivan, J.A. den Boon, P. Ahlquist, A positive-strand RNA virus replication complex parallels form and function of retrovirus capsids, Mol. Cell 9 (2002) 505–514.
[15] A. Noueiry, P. Ahlquist, Brome mosaic virus RNA replication: revealing the role of the host in RNA virus replication, Annu. Rev. Phytopathol. 41 (2003) 77–98.