Applied Mathematics Letters 40 (2015) 84–89

Contents lists available at ScienceDirect

Applied Mathematics Letters

journal homepage: www.elsevier.com/locate/aml

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

ARTICLE INFO

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

ABSTRACT

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.

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/

Table 1
BMV model results for naive time points and naive inputs (A), optimized time points and naive inputs for D-, E-, and SE-optimal designs (B-D through B-SE), or naive time points and optimized inputs for D-, E-, and SE-optimal designs (C-D through C-SE). NSE = normalized standard error.

Parameter	r_y	d_y	d_z	m	s
Estimate	31.641	0.7562	0.3139	0.5557	1.2374
NSE (A)	0.2223	0.6651	0.1947	2.9583	0.4318
95% CI (A)	(17.8575,45.4245)	(−0.22964,1.742)	(0.19414,0.43366)	(−2.6663,3.7777)	(0.19025,2.2846)
NSE (B-D)	0.1632	0.5402	0.1444	2.0333	0.3385
95% CI (B-D)	(21.52,41.762)	(−0.0445,1.5569)	(0.22508,0.40272)	(−1.6589,2.7703)	(0.41635,2.0584)
NSE (B-E)	0.1526	0.5152	0.1356	1.9022	0.3218
95% CI (B-E)	(22.1797,41.1023)	(−0.0074757,1.5199)	(0.2305,0.3973)	(−1.5161,2.6275)	(0.45694,2.0179)
NSE (B-SE)	0.1482	0.5032	0.1329	1.8226	0.3256
95% CI (B-SE)	(22.4505,40.8315)	(0.010449,1.502)	(0.23214,0.39566)	(−1.4294,2.5408)	(0.44772,2.0271)
NSE (C-D)	0.0744	0.0820	0.0940	0.3082	0.0454
95% CI (C-D)	(27.0296,36.2524)	(0.63472,0.87768)	(0.25607,0.37173)	(0.22,0.8914)	(1.1273,1.3475)
NSE (C-E)	0.0519	0.1669	0.0587	0.3471	0.0770
95% CI (C-E)	(28.4206,34.8614)	(0.50884,1.0036)	(0.2778,0.35)	(0.17765,0.93375)	(1.0507,1.4241)
NSE (C-SE)	0.0607	0.0813	0.0643	0.2981	0.0530
95% CI (C-SE)	(27.8745,35.4075)	(0.63564,0.87676)	(0.27431,0.35349)	(0.23107,0.88033)	(1.1089,1.3659)

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

$$\begin{aligned} &\frac{d\vec{x}}{dt} = \vec{g}(t, \vec{x}(t; \vec{\theta}, b(t)), \vec{q}, b(t)), \quad t \in [t_0, t_f], \\ &\vec{x}(t_0, \vec{\theta}) = \vec{x}_0, \qquad \vec{f}(t, \vec{\theta}, b(t)) = C\vec{x}(t, \vec{\theta}, b(t)) \end{aligned} \tag{1}$$

where $\vec{x}(t, \vec{\theta})$ is the vector of state variables of the system generated using a parameter vector $\vec{\theta} = (\vec{x}_0, \vec{q}) \in \mathbb{R}^p$, $p = N + r$, that contains N initial values and r system parameters listed in $\vec{q}$. The map $\vec{g} : \mathbb{R}^{1+N+r} \to \mathbb{R}^N$ has the corresponding observation process $\vec{f}(t, \vec{\theta}, b(t)) = C\vec{x}(t, \vec{\theta}, b(t))$ 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 $\vec{b}$ of length H, with values in $\{0, 1\}$ that represent whether the experimental input is on or off in the time intervals $[t^b_{i-1}, t^b_i]$, $i = 1, \ldots, H$.

For a given member θ_k of the estimated parameter vector $\vec{\theta}$ the standard error (SE_k) is computed by standard methods from asymptotic theory. For Tables 1 and 2, the normalized standard error (NSE) is defined as $\frac{\theta_k}{SE_k}$; 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 $B = \mathbb{Z}_2^H$, the set of binary vectors $\vec{b}$ 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

$$\mathcal{F}(\mathcal{P}_1, \mathcal{P}_2, \vec{\theta}_0) = \int_{t_0}^{t_f} \int_{\mathbb{Z}_2^H} \nabla^T_{\vec{\theta}_0} \vec{f}(t, \vec{\theta}_0, b(t)) \left(V_0^{-1}(t)\right) \nabla_{\vec{\theta}_0} \vec{f}(t, \vec{\theta}_0, b(t)) d\mathcal{P}_2(b) d\mathcal{P}_1(t). \tag{2}$$