Discrimination Between Competing Model Structures of Biological Systems in the Presence of Population Heterogeneity

MARC MARTIN-CASAS AND ALI MESBAH
Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA 94720 USA
CORRESPONDING AUTHOR: ALI MESBAH (mesbah@berkeley.edu)
The open-access publication was supported in part by the Berkeley Research Impact Initiative (BRII). The work of M. Martin-Casas was supported by La Caixa Fellowship Program.

ABSTRACT Computational models are useful for quantitative elucidation of the dynamical behavior of biological systems. Oftentimes, several competing models (i.e., hypotheses) are proposed to describe the underlying molecular mechanisms of a biological system. Selecting the most representative model is imperative for obtaining meaningful quantitative insights into the dynamics of the system of interest. However, the discrimination between competing models poses a significant challenge due to heterogeneity that is intrinsic to biological systems. This letter demonstrates the effectiveness of a probabilistic approach to optimal experiment design for model discrimination in the presence of time-invariant cell-to-cell differences within a cell population. The JAK2/STAT5 signaling pathway, which is involved in proliferation and differentiation of hematopoietic stem cells, is used as a case study.

INDEX TERMS Design of experiments, model discrimination, population heterogeneity.

I. INTRODUCTION

Computational models can play a central role in unraveling the fundamental mechanisms of biological systems, as they facilitate systematic design of experiments and dynamical analysis of complex networks of biochemical reactions [1]. The predictive quality of computational models critically hinges on selection of adequate descriptions of molecular mechanisms (i.e., model structures) that characterize the system dynamics. Multiple competing model structures, which correspond to different hypotheses for molecular mechanisms, typically exist for describing the dynamics of the same cellular processes. Optimal experiment design (OED) tools for model discrimination can assist in the invalidation of model structures that cannot adequately describe experimental data (see [2] and [3]). However, a key challenge in OED for biological systems arises from the presence of heterogeneity within a given cell population, i.e., cellular traits may present a certain degree of variability among individual cells in a population, i.e., cell traits may differ within a cell population [4]. The heterogeneity intrinsic to biological systems can significantly increase the complexity of model discrimination.

Numerous studies have elucidated the critical role of population heterogeneity in major biological processes such as embryogenesis, tumorigenesis, and survival [5]–[7]. Biological sources of cell population variability can be broadly classified into genetic and non-genetic. Genetic sources arise from spontaneous or induced mutation of the genetic material within a cell population. On the other hand, generation of variability in isogenic cell populations has revealed the importance of nongenetic factors that give rise to phenotypic heterogeneity [4]. Nongenetic heterogeneity mainly results from the following:

1) Temporal noise, that is, fluctuations of traits in individual cells due to intrinsic sources such as the randomness associated with gene expression and interactions at the molecular level;

2) Population noise, which originates from differences in time-invariant traits among individual cells within a cell population [6].

In this letter, modeling of the population variability exclusively focuses on time-invariant cell-to-cell differences in cellular traits within a population, which give rise to population noise.

The goal of this letter is to demonstrate the effectiveness of using a probabilistic OED framework [8], [9] for model discrimination in the presence of population heterogeneity. Probabilistic OED allows for designing system input(s) that can discriminate between distributions of cellular outputs (within a cell population) predicted by the competing models. Thus, upon application of the optimal input(s) to the biological system, invalidation of inadequate model structures can be achieved. In this letter, probabilistic OED is performed for the erythropoietin (Epo)-induced JAK2/STAT5 signaling pathway, for which several competing model structures can describe experimental data equally well under nominal stimulation levels [10]. The population heterogeneity that arises from population noise is modeled by time-invariant...
probabilistic distribution of kinetic parameters of the underlying biochemical reactions of the signaling pathway.

II. MODELING POPULATION NOISE IN THE JAK2/STAT5 SIGNALING PATHWAY

The family of JAK/STAT signaling pathways is ubiquitous in mammalian cells and is of relevance to the fields of stem cell and cancer research—its activation leads to modulation of expression of genes involved in growth, differentiation, migration, apoptosis, and other vital cellular processes [11]–[15]. JAK/STAT signaling can be stimulated with a large range of hormones, cytokines, and growth factors [16]. In mammals, four different JAKs and seven different STATs exist [16]. This letter focuses on the erythropoietin-stimulated pathway involving JAK2 and STAT5 (i.e., JAK2/STAT5 pathway), which is critical for growth and differentiation of hematopoietic progenitor cells [17], [18]. The key components of the signaling system include: 1) Epo receptor (EpoR), which is a transmembrane receptor that interacts with the extracellular pathway ligand; 2) Janus Kinase 2 (JAK2), which interacts with the cytosolic domain of EpoR; and 3) signal transducer and activator of transcription 5 (STAT5), which can be in monomeric or dimeric form.

A schematic of the JAK2/STAT5 signaling system is shown in Fig. 1. EpoR subunits undergo multimerization upon stimulation by Epo [16]. The cytoplasmic domain of each EpoR monomer interacts with a JAK2 tyrosine kinase. When two JAK2 proteins transphosphorylate each other, they become functionally active and induce phosphorylation of EpoR and other downstream signaling biomolecules, including STAT5 monomers. Phosphorylated STAT5 monomers dimerize in the cytoplasm and translocate into the nucleus, where they act as a transcription factor for upregulation of target genes. The dynamics of the JAK2/STAT5 signaling system are typically described by the four forms of the STAT5 molecules in the pathway [10]: unphosphorylated cytoplasmic STAT5 monomer \( x_1 \), phosphorylated cytoplasmic STAT5 monomer \( x_2 \), cytoplasmic STAT5 dimer \( x_3 \), and the transcriptionally active nuclear STAT5 dimer \( x_4 \). The input to the pathway, \( u \), consists of the level of phosphorylation of EpoR, which can be stimulated by extracellular Epo.

Swameye et al. [10] have reported several model structures for the JAK2/STAT5 signaling pathway in murine B cells. This letter considers two competing models of the signaling pathway (denoted by superscripts [1] and [2]) that could describe the population average experimental results reported in [10] with comparable statistical significance. Model 1, which assumes nucleocytoplasmic shuttling of STAT5 in the cell, takes the form

\[
\begin{align*}
\dot{x}_1^{[1]} &= -k_1 x_1^{[1]} u(t) + 2 k_4 x_3^{[1]} (t - \tau), \quad x_1^{[1]}(0) = x_{1,0} \\
\dot{x}_2^{[1]} &= -k_2 (x_2^{[1]})^2 + k_1 x_1^{[1]} u(t), \quad x_2^{[1]}(0) = x_{2,0} \\
\dot{x}_3^{[1]} &= -k_3 x_3^{[1]} + \frac{1}{2} k_2 (x_2^{[1]})^2, \quad x_3^{[1]}(0) = x_{3,0} \\
\dot{x}_4^{[1]} &= k_3 x_3^{[1]} - k_4 x_3^{[1]} (t - \tau), \quad x_4^{[1]}(0) = x_{4,0}.
\end{align*}
\]

FIGURE 1. Depiction of the JAK2/STAT5 signaling pathway.

Model 2, which also assumes nucleocytoplasmic shuttling while accounting for a bidirectional kinetic flow in the dimerization of STAT, is described by

\[
\begin{align*}
\dot{x}_1^{[2]} &= -k_1 x_1^{[2]} u(t) + 2 k_4 x_3^{[2]} (t - \tau), \quad x_1^{[2]}(0) = x_{1,0} \\
\dot{x}_2^{[2]} &= -k_2 (x_2^{[2]})^2 + k_1 x_1^{[2]} u(t) + 2 k_2 x_3^{[2]}, \quad x_2^{[2]}(0) = x_{2,0} \\
\dot{x}_3^{[2]} &= -k_3 x_3^{[2]} + \frac{1}{2} k_2 (x_2^{[2]})^2 - k_2 x_3^{[2]}, \quad x_3^{[2]}(0) = x_{3,0} \\
\dot{x}_4^{[2]} &= k_3 x_3^{[2]} - k_4 x_3^{[2]} (t - \tau), \quad x_4^{[2]}(0) = x_{4,0}.
\end{align*}
\]

In (1) and (2), \( k_1 \) and \( k_2 \) are the rates of phosphorylation and dimerization of STAT, respectively, \( k_2 \) is the rate of spontaneous disintegration of STAT dimers into monomers, \( k_3 \) and \( k_4 \) are the rates of transport in and out of the nucleus, respectively, and \( \tau \) is the residence time of STAT dimers in the nucleus. The (measurable) outputs of the pathway are phosphorylated STAT5 and total STAT5 in the cytoplasm, which are denoted by \( y_1 = k_5 (x_2 + 2 x_3) \) and \( y_2 = k_6 (x_1 + x_2 + 2 x_3) \), respectively, with \( k_5 \) and \( k_6 \) being scaling parameters. In a population of cells, cell-to-cell differences in kinetic rates or initial conditions can give rise to population noise. To capture population noise in competing models (1) and (2), the kinetic rates and time delay parameter are assumed to take beta distributions across the population of cells. The bounds of the beta distributions were adopted from the confidence intervals provided in [10]. The time-invariant differences in the kinetic and time delay characteristics of individual cells in a population will give rise to noise in constituents of the signaling pathway (and thus in the outputs \( y_1 \) and \( y_2 \)). The impact of each kinetic and time delay mechanism on heterogeneity of pathway outputs is likely to be different, reflecting their unequal contributions to the pathway dynamics. An analysis
of the impact of variability in the kinetic rates and time delay parameter on population noise can provide insight into molecular mechanisms of generation of population noise.

For illustrative purposes, the effect of probabilistic variability in $k_3$ on population noise is presented in Fig. 2, which shows snapshots of population noise in the output $y_1$ at the time instant 60 min (predicted by Model 1). The distributions of $y_1$ are constructed based on propagating the probability distributions of the kinetic rates and time delay parameter through the pathway dynamics using Monte Carlo simulations. In Fig. 2(a), $k_3$ is assigned a population-average value, whereas in Fig. 2(b) and (c), $k_3$ is distributed, respectively, with the same and triple the standard deviation reported in [10]. It is evident that the variability of $k_3$ across individual cells of a population affects the population noise of the output $y_1$.

III. MODEL DISCRIMINATION IN THE PRESENCE OF POPULATION NOISE

This letter seeks to discriminate between predictions of competing models (1) and (2) in the presence of population noise. Fig. 3 shows the dynamics of cytoplasmic phosphorylated STAT5 ($y_1$) predicted by the competing models upon EpoR stimulation with a nominal level of extracellular erythropoietin. The predictions of population average $y_1$ by the competing models (depicted by continuous profiles) are very similar, which are consistent with the results reported in [10]. More importantly, the predicted distributions of $y_1$ (due to population noise), which can be seen from distribution snapshots in Fig. 3 (at times 30, 45, and 60 min), overlap significantly. This makes the discrimination between predictions of models (1) and (2) impractical under the nominal stimulation level.

A probabilistic OED approach [8], [9] is employed for discriminating between competing models (1) and (2) in the presence of population noise. The distinct feature of the probabilistic OED approach lies in the ability to systematically account for population noise generation in the signaling pathway when the optimal input for model discrimination is designed. The probabilistic OED approach aims to separate the predicted distributions of pathway outputs within a population of cells. If the predicted distributions of one of the outputs are separated at least at one time instant, the competing models can be discriminated. Fig. 4(a) shows the evolution of the distribution of $y_1$ when the optimal input designed by probabilistic OED is applied to competing models (1) and (2) under 1000 realizations of the sources of population noise. The optimal input allows for discriminating between the predicted distributions of $y_1$ within a cell population at 60 min.

The performance of the optimal input designed by the probabilistic OED is compared to that of the standard OED (see [19]), which disregards the impact of time-invariant cell-to-cell differences within a population of cells. The optimal input designed by the standard OED is able to separate the population average predictions of the competing models [compare Fig. 4(b) with Fig. 3]. However, it fails to effectively discriminate between the predicted distributions of $y_1$ across the cell population.

IV. DISCUSSION

The use of computational models in combination with state-of-the-art measurement techniques can significantly improve our understanding of cellular processes that are prone to cell-to-cell variability. The above-discussed OED approaches are intended to complement experimental efforts, in particular when collecting rich experimental data is limited by various technical constraints and/or availability of resources. OED tools are invaluable for reducing the experimental effort required for elucidating of competing hypotheses of molecular mechanisms in biological systems. Accounting for the presence of cell-to-cell variability in a population of cells can be critical to achieving accurate quantitative representations of the underlying dynamics of a cell population. Traditional quantitative methods (such as western blotting and qPCR) typically yield end point and population average information and, consequently, can mask heterogeneity within a population [20]. On the other hand, real-time single-cell resolution techniques can surface cell-to-cell differences and, if high throughput, can provide a complete picture of population heterogeneity [20]. The above-discussed probabilistic OED approach enables discrimination between
competing models regardless of the employed measurement technique (population-average or single-cell measurement). A comparison of the results of standard and probabilistic OED approaches reveals that accounting for population heterogeneity allows for more effective experiment designs to achieve model discrimination with high confidence. In the presented case study, the probabilistic OED not only achieved a better discrimination of the population average behavior of the outputs compared with the standard OED, but also led to separation of the distributions of outputs of the competing models.

V. METHODS: PROBABILISTIC OED

The probabilistic OED problem is formulated as a nonlinear optimization problem that aims to maximize the dissimilarity between probability distributions of outputs of competing models [8], [9]. The optimization problem is subject to input and state constraints. The Kolmogorov distance [21] is used as a metric to quantify (dis)similarity between outputs of competing models. The generalized polynomial chaos (gPC) framework is used for propagation of population noise over pathway dynamics [22]. In gPC, each (probabilistic) system state is approximated by an expansion of orthogonal polynomial basis functions, which are defined based on the known probability distribution of kinetic parameters. The statistical moments of states can be efficiently computed from the coefficients of PC expansions or, alternatively, PC expansions can be used as a surrogate for the nonlinear system model to efficiently perform Monte Carlo simulations.

REFERENCES